University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 370

Stochastic Choice and Preference Reversals

Carlos Alós-Ferrer, Johannes Buckenmaier and Michele Garagnani

Revised version, July 2021

Stochastic Choice and Preference Reversals*

Carlos Alós-Ferrer†[1], Johannes Buckenmaier[1], and Michele Garagnani[1]

[1]Department of Economics, University of Zurich.

This Version: July 2021

Abstract

Preferences over risky alternatives can be elicited by different methods, including direct pairwise choices and willingness-to-accept valuations. The results are frequently at odds, casting doubts on the foundations of economics. We develop a stochastic choice model predicting when inconsistencies across elicitation methods should occur, the type of anomalies to be expected, what determines their magnitude, and whether they uncover a bias or not. While some anomalies can be traced back to individual biases, other apparent anomalies can occur in the absence of any actual behavioral bias, as a consequence of regularities in stochastic choice, risk attitudes, and experimental design. The model delivers new predictions that are confirmed in five experiments on the classical preference reversal phenomenon. Our novel empirical approach relies on utilities estimated out of sample, which allow us to test the model and also show that the bias in willingness-to-accept valuations is limited to long shots.

JEL Classification: D01 · D81 · D91

Keywords: Stochastic choice · Preference elicitation · Preference reversals · Behavioral biases · Lottery choice

*We are grateful to Miguel Ballester, Sudeep Bhatia, Jordi Brandts, Ernst Fehr, Graham Loomes, Ganna Pogrebna, and Roberto Weber for helpful comments and discussions. Financial support from the German Research Foundation (DFG) through project Al-1169/4 (Research Unit "Psychoeconomics," FOR 1882), and from the Swiss National Science Foundation (SNF) under project nr. 100014_179009 is gratefully acknowledged.

†Corresponding author. Zurich Center for Neuroeconomics (ZNE), Department of Economics, University of Zurich (Switzerland). Blümlisalpstrasse 10, CH-8006 Zurich. E.mail: carlos.alos-ferrer@econ.uzh.ch

1 Introduction

A preference elicitation anomaly occurs when two different but theoretically-equivalent preference elicitation methods contradict each other. For example, a decision maker might reveal a higher willingness to pay for option A than for option B, but then actually choose B over A when given opportunity. The most prominent example in the domain of risky choice is the classical preference reversal phenomenon (Lichtenstein and Slovic, 1971; Grether and Plott, 1979; Tversky and Thaler, 1990), where monetary valuations of gambles contradict risky choices. This anomaly has received enormous attention in economics (e.g., Holt, 1986; Karni and Safra, 1987; Tversky et al., 1990; Cubitt et al., 2004; Schmidt and Hey, 2004; Butler and Loomes, 2007), but it is just one of many. Inconsistencies between preference elicitation methods abound, including reversals between pricing and rating (Schkade and Johnson, 1989) and between certainty and probability equivalents (Hershey and Schoemaker, 1985; Johnson and Schkade, 1989; Delquié, 1993; Collins and James, 2015). They occur in multiple domains, ranging from health utility measurements (e.g. Stalmeier et al., 1997; Bleichrodt and Pinto Prades, 2009; Oliver, 2013; Attema and Brouwer, 2013) to decision-making under ambiguity (Maafi, 2011; Trautmann et al., 2011). The inconsistencies are robust, systematic, and highly relevant for economic analysis, because individual and societal preferences are often estimated on the basis of monetary valuations or similar constructs (see, e.g., Bateman et al., 2002 for a detailed discussion). Thus, if such measurements contradict actual choices, welfare economics and most of normative economics would be on shaky grounds. Moreover, discrepancies between elicitation methods are fundamentally at odds not only with Expected Utility Theory, but with any preference-based theory of decisions under risk assuming that decision makers' preferences can be represented by a stable utility function, including Cumulative Prospect Theory and Rank-Dependent Utility.

In the present work, we develop and test a stochastic choice model that provides a unified account of preference elicitation anomalies in risky choice, while also deriving new testable predictions. This is accomplished by incorporating received insights on the structure of errors from the stochastic choice literature, which also allows us to directly model biases in choice and valuation. Specifically, we postulate a monotonic relation between error rates and 'strength of preference,' captured by differences in certainty equivalents. In the absence of a systematic bias, this implies that error rates should be larger when differences in certainty equivalents are small.[1]

The model considers settings where preferences within pairs of alternatives are elicited according to two different methods. To fix ideas, suppose they correspond to direct choices and some kind of indirect evaluations (but the model encompasses any comparison across methods). Alternatives contain options of two well-differentiated types (e.g.,

[1]This property, which is a standard assumption in random utility models (McFadden, 2001), arises from long-standing insights from psychophysics (Dashiell, 1937; Moyer and Landauer, 1967) and has been recently demonstrated in the domain of decisions under risk (Alós-Ferrer and Garagnani, 2018).

long-shots and moderate lotteries). Since choices and evaluations are stochastic, a number of discrepancies between the methods (reversals) is natural. A preference elicitation anomaly arises if an asymmetry is observed, namely if the proportion of reversals of one type systematically exceeds the proportion of reversals of the opposite type. For instance, the classical preference reversal phenomenon reduces to the observation that decision makers frequently choose moderate lotteries over long shots of similar expected value, but then reveal a higher monetary valuation for the long shots, while the opposite reversals are rare. Hence the rate for the first kind of reversals (proportion of pairs where the reversal occurs over all the pairs where the moderate lottery is chosen) is much higher than the analogous rate for the second kind (Grether and Plott, 1979; Tversky and Thaler, 1990). Such asymmetries have universally been taken as evidence that reversals cannot be due just to random errors arising from stochasticity in choices, elicited valuations, or both (e.g., Schmidt and Hey, 2004; Loomes, 2005).

The most important insight arising from the model is that the overall proportion of choices and evaluations in favor of one of the types of alternatives is a crucial determinant of the rates of reversals. Consequently, a given experiment can itself be biased toward one kind of alternatives, in the sense that they are chosen more frequently on average. This can happen for reasons which are completely orthogonal to any behavioral bias of the decision makers. For instance, the particular attitudes toward risk in the experimental sample, or the selection of lottery pairs to fulfill certain criteria, will generally result in a biased experiment. Strikingly, the model shows that preference elicitation anomalies can occur *even in the absence of any behavioral biases*, that is, apparent, systematic anomalies where one reversal rate is predictably larger than the opposite rate can be created experimentally out of thin air even though both elicitation methods are equivalent (but noisy).

The model also predicts other anomalies, as the one underlying the classical preference reversal phenomenon, as a consequence of strength of preference and a behavioral bias in an evaluation method. We obtain comparative-statics results showing that a stronger behavioral bias in evaluations exacerbates the anomalies, but the degree to which an experiment is biased (toward the option chosen in the most frequent reversals) has the opposite effect. In particular, the implicit assumption in the literature that in the absence of an evaluation bias one should expect comparable rates of reversals is incorrect. This assumption is only justified if additionally the experiment itself is also unbiased. For example, if an experiment on the classical preference reversal phenomenon is biased toward moderate lotteries, then in the absence of a behavioral bias the model would predict the *opposite* of the preference reversal phenomenon.

To test the model's predictions, and also to motivate and test the validity of the model's assumptions, we conducted five experiments (total $N = 503$) focused on the classical preference reversal phenomenon, relying on a novel empirical approach which includes utility estimations out of sample. The design provides empirical evidence not previously available in preference reversal experiments. We find evidence for strength-

of-preference effects both for actual choices and for imputed choices derived from the comparison of elicited valuations. Our empirical tests then include the preference reversal phenomenon, the reversal of the phenomenon (which arises in the absence of any behavioral bias), and novel predictions on the causal effects of a behavioral bias in valuations and of biased experiments. Here, the estimated preferences become essential: On the one hand, they serve as a natural scale to study the regularities in stochastic choice that underlie the model. On the other hand, they allow us to test novel predictions of the model, which link reversals to individual risk attitudes and to a bias in valuations seen as a deviation from the own preferences.

Our experimental evidence also provides further insights. First, there is no general overpricing bias in elicited monetary valuations. Instead, monetary valuations for moderate lotteries are very accurate, which is at odds with the generalized impression in the literature. In contrast, subjects dramatically overstate their monetary valuations for long shots. That is, there is a systematic bias in subjects' monetary valuations but it is confined to one type of lotteries, namely long shots. This specific bias in monetary valuations is shown to cause the classical preference reversal phenomenon. Strikingly, we show that most preference reversal experiments in the literature were actually set up in such a way that if there was no bias at all in monetary valuations, then the reversal of the preference reversal phenomenon should occur. Because the expected values of both lotteries within a pair are usually very similar in preference reversal experiments (but long shots are riskier) and most subjects are risk averse, experiments in this domain are often biased toward the moderate lotteries, in the sense that the latter are chosen more frequently on average. This yields two important insights: First, the literature has underestimated the extent of the preference reversal phenomenon for half a century, by comparing it to an incorrect default (the equality of reversal rates). Second, the reversal of the preference reversal phenomenon, which had been observed and considered a puzzle, is *not* the result of another bias in evaluations but rather the direct consequence of stochastic choice and risk aversion (resulting in a biased experiment). Additionally, our results allow us to encompass, clarify, and organize previous empirical findings which were hard to interpret up to now.

Robustness analyses confirm that our results do *not* depend on the specific features of our experimental design or of the estimation procedure. First, they do not obtain simply because preferences estimated from choices (although out-of-sample) reflect other choices more accurately than valuation decisions. In fact, the above results remain robust when alternative utility functions are estimated from an independent set of imputed choices derived from the comparison of elicited valuations (stated prices). The reason is simply that there is no systematic bias between choices and monetary valuations in general, but rather a bias in the valuation of long shots only. Second, the results obtain for different utility functions and different estimation procedures. Third, the results obtain independently of whether valuations are incentivized via an intuitive ordinal payoff method or

a Becker-DeGroot-Marschak procedure. Fourth, the results are qualitatively unchanged when willingness-to-pay valuations are used instead of willingness-to-accept valuations.

The paper is structured as follows. Section 2 presents our stochastic choice model and provides a unified account of preference elicitation anomalies. Section 3 briefly reviews the classical preference reversal phenomenon and paves the way for our empirical application. Section 4 describes our experimental design and the utility estimation procedure. Section 5 presents our results for preference reversal experiments with unbiased as well as biased evaluations and discusses how previous empirical findings are accounted for in light of our results. Section 6 briefly summarizes a number of robustness analyses. Section 7 concludes. The Supplementary Materials (Online Appendix) contain additional details.

2 A Stochastic Choice Model for Preference Elicitation Anomalies

In this section we develop a stochastic choice model for the analysis of preference elicitation anomalies. The model allows us to derive novel comparative static predictions, that we test and confirm with our data in later sections.

Consider an experiment where a decision maker (DM) is asked to express her preferences for a set D of K lottery pairs $D = \{(\mathrm{P}_1, \$_1), \ldots, (\mathrm{P}_K, \$_K)\}$, using two different elicitation procedures to which we refer as "choice" and "evaluation" for simplicity (but can actually be arbitrary elicitation methods). Each pair is made out of lotteries of two well-differentiated types, a P-bet lottery P_k and a \$-bet lottery $\$_k$. For the formal model, these categories are abstract, but we choose the names to make the application to the preference reversal phenomenon in later sections transparent. For our purposes, a DM can be characterized by a *stochastic choice function* ρ_c and a *stochastic evaluation function* ρ_v such that for any lottery pair $(\mathrm{P}, \$)$ the probability that the DM chooses P over \$ is given by $\rho_c(\mathrm{P}, \$)$, and the probability that P is evaluated higher than \$ is given by $\rho_v(\mathrm{P}, \$)$.

A preference reversal occurs for a pair $(\mathrm{P}_k, \$_k)$ if the preferences elicited from choices and evaluations are inconsistent. There are two types of reversals. In a *standard reversal (SR)*, the lottery P_k is chosen over $\$_k$ but $\$_k$ is evaluated higher than P_k. In a *non-standard reversal (NR)* the lottery $\$_k$ is chosen over P_k but P_k is evaluated higher than $\$_k$. For a DM the likelihood to observe a standard reversal for a pair $(\mathrm{P}_k, \$_k)$ is $\rho_c(\mathrm{P}_k, \$_k)(1 - \rho_v(\mathrm{P}_k, \$_k))$, whereas the likelihood to observe a non-standard reversal is $(1 - \rho_c(\mathrm{P}_k, \$_k))\rho_v(\mathrm{P}_k, \$_k)$. The rate of standard (resp. non-standard) reversals in experiments is computed as the number of standard (resp. non-standard) reversals divided

by the number of P-choices (resp. \$-choices). For a DM characterized by (ρ_c, ρ_v) the expected rate of standard reversals in experiment D is

$$SR(D, \rho_c, \rho_v) = \sum_{k=1}^{K} \frac{\rho_c(\mathrm{P}_k, \$_k)}{\sum_{\ell=1}^{K} \rho_c(\mathrm{P}_\ell, \$_\ell)} (1 - \rho_v(\mathrm{P}_k, \$_k)) \tag{1}$$

and the expected rate of non-standard reversals in experiment D is

$$NR(D, \rho_c, \rho_v) = \sum_{k=1}^{K} \frac{(1 - \rho_c(\mathrm{P}_k, \$_k))}{\sum_{\ell=1}^{K} (1 - \rho_c(\mathrm{P}_\ell, \$_\ell))} \rho_v(\mathrm{P}_k, \$_k). \tag{2}$$

Now, suppose that in addition to (ρ_c, ρ_v) the DM is endowed with a utility function u. For a pair $(\mathrm{P}, \$)$, we denote the difference in certainty equivalents (CE) by

$$\Delta(\mathrm{P}, \$) = u^{-1}(EU(\mathrm{P})) - u^{-1}(EU(\$))$$

where $EU(P)$ and $EU(\$)$ denote the respective expected utilities. The idea is that this difference may serve as an individual measure of preference strength with P being chosen more frequently and evaluated higher than \$ more often for larger values of $\Delta(\mathrm{P}, \$)$. Our stochastic choice model postulates that the functions ρ_c and ρ_v can be written as monotonically increasing functions of CE differences. That is, we assume that the propensity to choose or evaluate a P-bet over a \$-bet is increasing in the difference between their CEs, $\Delta(\mathrm{P}, \$)$. We say that a DM exhibits *strength-of-preference* (SoP) effects if ρ_c and ρ_v can be written as increasing functions of $\Delta(\mathrm{P}, \$)$. Thus, a DM exhibiting SoP effects is characterized by (ρ_c, ρ_v, u) and, hence, the expected reversal rates (1) and (2) can be written as functions of the CE differences $\Delta_k = \Delta(\mathrm{P}_k, \$_k)$,

$$SR(D, \rho_c, \rho_v, u) = \sum_{k=1}^{K} \frac{\rho_c(\Delta_k)}{\sum_{\ell=1}^{K} \rho_c(\Delta_\ell)} (1 - \rho_v(\Delta_k)) \tag{3}$$

and

$$NR(D, \rho_c, \rho_v, u) = \sum_{k=1}^{K} \frac{(1 - \rho_c(\Delta_k))}{\sum_{\ell=1}^{K} (1 - \rho_c(\Delta_\ell))} \rho_v(\Delta_k), \tag{4}$$

respectively.

We say that a DM *is unbiased* (in evaluations) if $\rho_v(\Delta) = \rho_c(\Delta)$ for all Δ, that is, there is no bias in evaluations relative to choice. On the other hand, we say that a DM *exhibits a \$-bias (P-bias)* in evaluations if $\rho_v(\Delta) < \rho_c(\Delta)$ ($\rho_v(\Delta) > \rho_c(\Delta)$) for all Δ. Of course reversals of both types will still occur independently of whether the DM is biased or not due to the fact that both choices and evaluations are stochastic. If a DM is unbiased, it follows immediately that the proportion of P-choices for a set of lotteries D, denoted by $\pi_c(D) = \frac{1}{K} \sum_{k=1}^{K} \rho_c(\Delta_k)$, equals the proportion of evaluations in favor of P, denoted by $\pi_v(D) = \frac{1}{K} \sum_{k=1}^{K} \rho_v(\Delta_k)$. However, whether P is chosen more frequently than \$ or vice versa depends on both the individual characteristics of the DM, that is,

ρ_c and u, and the set of lotteries D used in the experiment. We say that an experiment D is *biased toward P-bets (\$-bets)* for a DM if $\pi_c(D) > \frac{1}{2}$ ($\pi_c(D) < \frac{1}{2}$), that is, P-bets (\$-bets) are chosen more frequently. Note that an experiment might be biased although there is no behavioral bias at the DM level.

It is now easy to show that there is a direct link between the ratio of the standard and non-standard reversal rates and the quantities discussed above. Specifically, using (3) and (4) we obtain that

$$\begin{aligned}\frac{SR(D,\rho_c,\rho_v,u)}{NR(D,\rho_c,\rho_v,u)} &= \frac{\left(\sum_{\ell=1}^{K}(1-\rho_c(\Delta_\ell))\right)\cdot\sum_{k=1}^{K}\rho_c(\Delta_k)(1-\rho_v(\Delta_k))}{\left(\sum_{\ell=1}^{K}\rho_c(\Delta_\ell)\right)\cdot\sum_{k=1}^{K}(1-\rho_c(\Delta_k))\rho_v(\Delta_k)} \\ &= \frac{(1-\pi_c(D))}{\pi_c(D)}\left[\frac{\sum_{k=1}^{K}\rho_c(\Delta_k)-\rho_c(\Delta_k)\rho_v(\Delta_k))}{\sum_{k=1}^{K}\rho_v(\Delta_k)-\rho_c(\Delta_k)\rho_v(\Delta_k)}\right]\end{aligned}$$

which simplifies to

$$\frac{SR(D,\rho_c,\rho_v,u)}{NR(D,\rho_c,\rho_v,u)} = \frac{1-\pi_c(D)}{\pi_c(D)}\left[\frac{\pi_c(D)-\frac{1}{K}\sum_{k=1}^{K}\rho_c(\Delta_k)\rho_v(\Delta_k)}{\pi_v(D)-\frac{1}{K}\sum_{k=1}^{K}\rho_c(\Delta_k)\rho_v(\Delta_k)}\right] \tag{5}$$

We will consider two types of anomalies that refer to asymmetries between the rates of standard and non-standard reversals. We say that a DM exhibits a *type-1 anomaly* if $NR(D,\rho_c,\rho_v,u) > SR(D,\rho_c,\rho_v,u)$, that is, there is an asymmetry between reversal rates with more non-standard than standard reversals. Analogously, we say that a DM exhibits a *type-2 anomaly* if $SR(D,\rho_c,\rho_v,u) > NR(D,\rho_c,\rho_v,u)$, that is, there is an asymmetry between reversal rates with more standard than non-standard reversals. The classical preference reversal phenomenon corresponds to a type-2 anomaly.

2.1 Unbiased Evaluations

In this subsection, we consider the case of an evaluation method such that the DM has no bias in evaluation relative to choice ($\rho_v = \rho_c$). We show that the ratio between standard and non-standard reversals is decreasing in the proportion of P-choices, $\pi_c(D)$. This implies that a completely unbiased DM can produce both types of anomalies depending on $\pi_c(D)$, which is determined by the decision maker's risk attitude (as captured by u) and the properties of the lottery pairs used in experiment D. Specifically, an experiment that is biased toward P-bets ($\pi_c(D) > \frac{1}{2}$) leads to a type-1 anomaly, whereas an experiment that is biased toward \$-bets ($\pi_c(D) < \frac{1}{2}$) leads to a type-2 anomaly. Hence, both types of anomalies may arise as a consequence of individual characteristics of the DM (which are not a bias), SoP effects in stochastic choice, and the specifics of the experiment, even in the absence of any behavioral bias.

We first give the intuition of how our stochastic choice framework can explain type-1 and type-2 anomalies. This is illustrated in Figure 1. The key insight is that the comparison of standard and non-standard reversal rates hinges upon conditioning on

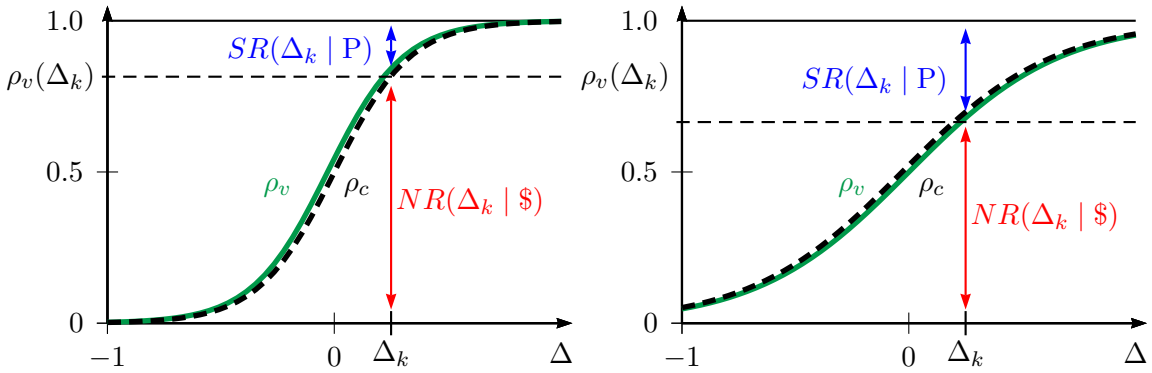


Figure 1: Unbiased evaluations.

the actual choice of either a P-bet or a \$-bet. For concreteness, fix a lottery pair $(\mathrm{P}_k, \$_k)$ with CE difference Δ_k. Then, conditional on P_k being chosen, the likelihood to observe a standard reversal for $(\mathrm{P}_k, \$_k)$ is simply the likelihood that $\$_k$ is evaluated higher than P_k, which is $SR(\Delta_k \mid \mathrm{P}) = 1 - \rho_v(\Delta_k)$. Analogously, conditional on a \$-bet being chosen the likelihood to observe a non-standard reversal is $NR(\Delta_k \mid \$) = \rho_v(\Delta_k)$.

Figure 1 illustrates two examples of an unbiased DM where $\rho_c(0)$ is exactly one half. Then, for any lottery pair with $\Delta_k = 0$, the likelihood to evaluate P_k above $\$_k$ is exactly 50%, and consequently we should expect similar rates of standard and non-standard reversals. However, if $\Delta_k > 0$, then the probability that P_k is evaluated above $\$_k$ exceeds 50% because ρ_v is increasing in Δ. Consequently, one should expect more non-standard than standard reversals for pairs with $\Delta_k > 0$, giving rise to a type-1 anomaly as the left panel of Figure 1 illustrates. Analogously, if $\Delta_k < 0$, then the probability that P_k is evaluated above $\$_k$ is below 50%. Thus, one should expect more standard than non-standard reversals for pairs with $\Delta_k < 0$ giving rise to a type-2 anomaly. That is, whether an anomaly is expected, and if so which one, depends on the CE differences $(\Delta_k)_{k=1}^K$, which in turn depend on the set of lotteries D used in the experiment and the DM's risk attitude.

To see this formally, note that if a DM is unbiased, then it follows that $\pi_c(D) = \pi_v(D)$. Equation (5) then simplifies dramatically to

$$\frac{SR(D, \rho_c, \rho_v, u)}{NR(D, \rho_c, \rho_v, u)} = \frac{1 - \pi_c(D)}{\pi_c(D)}.$$

The proof of the following result then immediately follows from this equation.

Proposition 1. *If a DM is unbiased, then the following statements hold.*

(i) An experiment D that is biased toward P-bets ($\pi_c(D) > \frac{1}{2}$) leads to a type-1 anomaly, that is, $SR(D, \rho_c, \rho_v, u) < NR(D, \rho_c, \rho_v, u)$.

(ii) An experiment D that is biased toward \$-bets ($\pi_c(D) < \frac{1}{2}$) leads to a type-2 anomaly, that is, $SR(D, \rho_c, \rho_v, u) > NR(D, \rho_c, \rho_v, u)$.

Proposition 1 shows that, even in the absence of a systematic difference between evaluations and choices, behavioral noise does *not* lead to equal reversal rates. On the contrary, whether a type-1 or a type-2 anomaly is expected depends on whether and toward which alternative the experiment is biased, which in turn is determined by individual characteristics of the DM and the set of lotteries used in the experiment.

The model also delivers a novel comparative static result. We say that the *bias in experiment D toward P-bets is stronger for DM i than for DM j* if $\pi_c^i(D) > \pi_c^j(D)$. Further, for a DM i we say that the *bias in experiment D_1 toward P-bets is stronger than in experiment D_2* if $\pi_c^i(D_1) > \pi_c^i(D_2)$. The next (straightforward) result shows that for unbiased DMs the relative proportion of standard to non-standard reversals decreases as an experiment becomes more biased toward P-bets.

Proposition 2. *The following statements hold.*

(a) Consider two unbiased DMs characterized by $(\rho_c^i, \rho_v^i, u^i)$ and $(\rho_c^j, \rho_v^j, u^j)$, respectively. If the bias in experiment D toward P-bets is stronger for i than for j, then

$$\frac{SR(D, \rho_c^i, \rho_v^i, u^i)}{NR(D, \rho_c^i, \rho_v^i, u^i)} < \frac{SR(D, \rho_c^j, \rho_v^j, u^j)}{NR(D, \rho_c^j, \rho_v^j, u^j)}.$$

(b) Consider an unbiased DM characterized by (ρ_c, ρ_v, u). If the bias in experiment D_1 toward P-bets is stronger than the bias in experiment D_2, then

$$\frac{SR(D_1, \rho_c, \rho_v, u)}{NR(D_1, \rho_c, \rho_v, u)} < \frac{SR(D_2, \rho_c, \rho_v, u)}{NR(D_2, \rho_c, \rho_v, u)}.$$

Hence, depending on whether an experiment exhibits type-1 or type-2 anomalies, a stronger bias (in the experiment) toward P-bets either exacerbates or dampens the anomaly, respectively. This will be useful for testing the model in later sections.

Another interesting observation is that also the level of consistency (noise) may affect the extent to which an experiment is biased toward P-bets. To see this, consider the right panel of Figure 1, which illustrates an example where both ρ_c and ρ_v are less consistent compared to the example in the left panel. Formally, we say that a stochastic choice (or evaluation) function ρ_1 is *less consistent* than ρ_2 if $\rho_1(\Delta) < \rho_2(\Delta)$ for $\Delta > 0$ and $\rho_1(\Delta) > \rho_2(\Delta)$ for $\Delta < 0$. It is easy to see that for a fixed $\Delta_k > 0$ ($\Delta_k < 0$) the proportion of P choices is smaller if choices and evaluations are less consistent. Hence, intuitively if most lottery pairs in an experiment D are such that $\Delta_k > 0$, then the proportion of P choices decreases as ρ_c becomes less consistent. In other words, depending on Δ_k the extent of the observed anomaly may decrease as the level of behavioral noise increases.

Summarizing, we have obtained an important and novel insight. Even with unbiased DMs, we should *expect* type-1 anomalies in any preference reversal experiment that is biased toward P-bets. Explaining this anomaly does not require any behavioral bias on the side of the DM, and its strength is monotonically related to the extent to which the experiment is biased toward P-bets as captured by $\pi_c(D)$.

2.2 Biased Evaluations

In this subsection, we consider the case of an evaluation method such that the DM has a systematic bias in evaluations relative to choice. The left panel of Figure 2 gives an illustrative example of a DM with a \$-bias in evaluations. The stochastic evaluation function is shifted downwards relative to the stochastic choice function. Intuitively, if there is a $\delta > 0$ such that $\rho_v(\delta) = \frac{1}{2}$, then δ captures the extent of the bias of evaluations toward the \$-bets, that is, δ is the premium that makes the decision maker (stochastically) indifferent between P and $\$+\delta$. Since $\rho_v(0)$ is below one half, for $\Delta_k = 0$, and also for pairs with positive but not-too-large differences in certainty equivalents, the likelihood to evaluate the P-bet above the \$-bet is smaller than 50%. As a result, we should expect more standard than non-standard reversals, i.e. a type-2 anomaly.

If a DM exhibits a \$-bias in evaluations, it follows that the proportion of P-choices exceeds the proportion of evaluations in favor of P, that is, $\pi_c(D) > \pi_v(D)$. Interestingly, however, a \$-bias in evaluations is not required for $\pi_c(D) > \pi_v(D)$ to obtain. For example, suppose that ρ_v is less consistent than ρ_c as the right panel of Figure 2 illustrates. Then, for a set of lottery pairs D with $\Delta_k > 0$ for all k, we have $\rho_c(\Delta_k) > \rho_v(\Delta_k)$ for all k, hence $\pi_c(D) > \pi_v(D)$. Consequently, if an experiment features enough decision problems with $\Delta_k > 0$, then $\pi_c(D) > \pi_v(D)$ may obtain merely because evaluations are less consistent than choices.

The essence of the effect of a DM exhibiting a \$-bias can be seen directly in equation (5). Suppose, as a thought experiment, that the experiment D was unbiased for the DM, in the sense that $\pi_c(D) = \frac{1}{2}$. It is then an immediate implication of equation (5) that a type-2 anomaly is predicted if the DM exhibits a \$-bias in evaluations. Formally, we obtain the following result.

Proposition 3. *Consider an unbiased experiment D in the sense that $\pi_c(D) = \frac{1}{2}$. If a DM exhibits a \$-bias in evaluations, then the DM displays a type-2 anomaly, that is, $SR(D, \rho_c, \rho_v, u) > NR(D, \rho_c, \rho_v, u)$.*

Proposition 4 below provides a second, novel comparative-statics prediction, namely *ceteris paribus* a stronger \$-bias in evaluations exacerbates the difference between the rates of standard and non-standard reversals. Formally, for two DMs i, j characterized by (ρ_c, ρ_v^i, u) and (ρ_c, ρ_v^j, u), we say that DM i exhibits a *stronger \$-bias (in evaluations)* than j in experiment D if $\rho_v^i(\Delta_k) < \rho_v^j(\Delta_k)$ for all k. Intuitively, a stronger \$-bias means that a DM is more likely to evaluate $\$_k$ higher than the P_k.

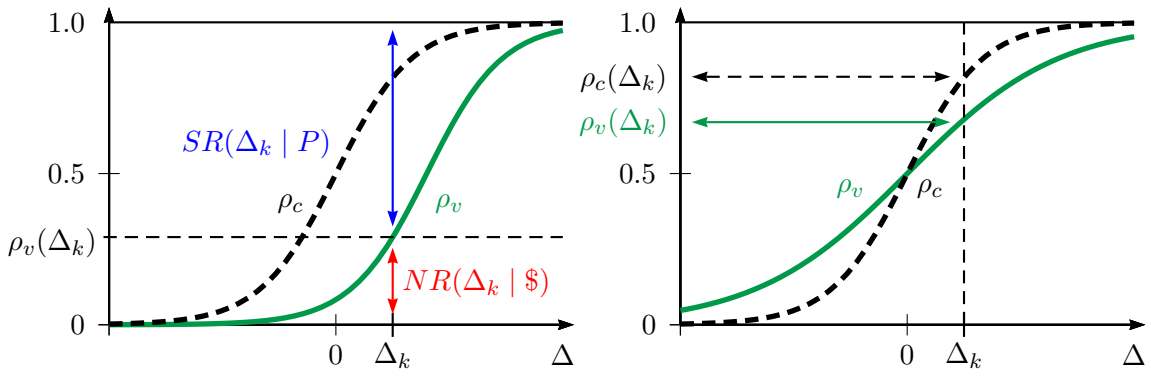


Figure 2: Stochastic evaluation functions and reversal rates.

Proposition 4. *Suppose two decision makers are characterized by* (ρ_c, ρ_v^i, u) *and* (ρ_c, ρ_v^j, u). *If DM* i *exhibits a stronger \$-bias than* j *in* D, *then*

$$SR(D, \rho_c, \rho_v^i, u) - NR(D, \rho_c, \rho_v^i, u) > SR(D, \rho_c, \rho_v^j, u) - NR(D, \rho_c, \rho_v^j, u)$$

for any stochastic choice function ρ_c.

Proof. Consider an arbitrary stochastic choice function ρ_c. We have

$$\begin{aligned} SR(D, \rho_c, \rho_v^i, u) - NR(D, \rho_c, \rho_v^i, u) &= \frac{\sum_{k=1}^K \rho_c(\Delta_k)(1 - \rho_v^i(\Delta_k))}{\sum_{\ell=1}^K \rho_c(\Delta_\ell)} - \frac{\sum_{k=1}^K (1 - \rho_c(\Delta_k))\rho_v^i(\Delta_k)}{\sum_{\ell=1}^K (1 - \rho_c(\Delta_\ell))} \\ &= \sum_{k=1}^K \frac{\rho_c(\Delta_k)}{\sum_{\ell=1}^K \rho_c(\Delta_\ell)} - \left(\frac{\rho_c(\Delta_k)}{\sum_{\ell=1}^K \rho_c(\Delta_\ell)} + \frac{(1 - \rho_c(\Delta_k))}{\sum_{\ell=1}^K (1 - \rho_c(\Delta_\ell))} \right) \rho_v^i(\Delta_k) \\ &> \sum_{k=1}^K \frac{\rho_c(\Delta_k)}{\sum_{\ell=1}^K \rho_c(\Delta_\ell)} - \left(\frac{\rho_c(\Delta_k)}{\sum_{\ell=1}^K \rho_c(\Delta_\ell)} + \frac{(1 - \rho_c(\Delta_k))}{\sum_{\ell=1}^K (1 - \rho_c(\Delta_\ell))} \right) \rho_v^j(\Delta_k) \\ &= SR(D, \rho_c, \rho_v^j, u) - NR(D, \rho_c, \rho_v^j, u) \end{aligned}$$

where the inequality follows because $\rho_v^i(\Delta_k) < \rho_v^j(\Delta_k)$ for all Δ_k since i exhibits a stronger \$-bias than j. This completes the proof. □

3 The Classical Preference Reversal Phenomenon

The classical preference reversal paradigm involves pairs of lotteries, typically consisting of a relatively safe lottery, called the P-bet ('P' for probability), and a riskier lottery offering a larger prize (a long shot), called the \$-bet. Individual preferences over such pairs are elicited independently in two ways, typically through pairwise choices and by eliciting valuations separately for each lottery, e.g. using stated minimal selling prices. The anomaly refers to the observation that decision makers often choose the comparatively-safe P-bet, but state a larger monetary valuation for the \$-bet than for the P-bet, which corresponds to a standard preference reversal in the terms of our model. This empirical

pattern is extremely robust (in the words of Butler and Loomes, 2007, "easy to produce, but much harder to explain;" see Seidl, 2002 for a comprehensive survey). Crucially, however, standard reversals occur much more frequently than non-standard reversals, in which \$-bets are chosen but P-bets receive a higher valuation.

Even more striking than the preference reversal phenomenon is the fact that, if the monetary valuation task is replaced by an ordinal ranking task, the anomaly is reversed. That is, instead of resulting in similar rates of standard and non-standard reversals, this alternative implementation results in a *reversal of the preference reversal phenomenon* (Casey, 1991; Bateman et al., 2007; Alós-Ferrer et al., 2016) where non-standard reversal rates, which are rather low in the original design, now exceed the standard ones. This is striking, because ranking tasks are conceptually closer to binary choices, and hence should avoid systematic differences across elicitation methods. This puzzling reversal of the phenomenon cannot be accounted for by any of the explanations of preference reversals previously proposed in the literature. For instance, Bateman et al. (2007) argued that ranking methods introduced "distorting effects of their own."

A large number of competing, partial explanations for the preference reversal phenomenon has been put forward over the years, including systematic violations of transitivity (Safra et al., 1990) or procedural invariance (Goldstein and Einhorn, 1987), among others. The prominence hypothesis (Tversky et al., 1988) generally attributes inconsistencies to choice errors, arguing that decision makers focus on a prominent attribute (e.g., the winning probability) and overweight it in choice tasks compared to evaluation tasks. The scale compatibility hypothesis (Tversky et al., 1990) attributes the phenomenon to errors in the evaluation method, arising because decision makers overweight attributes which naturally map onto the (monetary) evaluation scale leading to so-called overpricing. Accounts based on choice inconsistencies (Schmidt and Hey, 2004) argue that preference reversals occur because evaluation tasks are less natural and hence more noisy than choices, resulting in differences in error rates. However, starting with the seminal work of Tversky et al. (1990) it has been consistently shown that the overall phenomenon persists in experimental settings that control for these explanations (e.g. Pommerehne et al., 1982; Cubitt et al., 2004). While there is general agreement that several of the explanations given above influence preference reversals, no single account has been able to fully explain when and why the phenomenon should be expected, and which factors ameliorate or exacerbate it.

Our formal model applies directly to the classical preference reversal phenomenon (and its reversal). P-bets and \$-bets are lotteries of types P and \$, respectively. The choice and valuation methods correspond to the stochastic choice and valuation functions, ρ_c and ρ_v. The preference reversal phenomenon is a type-2 anomaly, while its reversal is a type-1 anomaly. All our results apply directly. Hence, the model explains both the preference reversal phenomenon and its reversal (and, in particular, that the latter is *not* due to any new bias or additional distorting effects), while also deriving new testable predictions.

Table 1: Summary of experiments

Experiment	N	Lab	Preference reversal experiment		Out-of-sample estimation based on		Incentivization of WTA/WTP
			Choice	Evaluation	choices	evaluations	
RANK1	95	Cologne	Yes	Ranking	Yes	No	-
RANK2	108	Zurich	Yes	Ranking	Yes	Yes (WTA)	BDM
WTA1	95	Cologne	Yes	WTA	Yes	No	OPM
WTA2	103	Zurich	Yes	WTA	Yes	Yes (WTA)	BDM
WTP	102	Zurich	Yes	WTP	Yes	Yes (WTP)	BDM

4 The Experiments

We have developed a stochastic choice model explaining type-1 and type-2 anomalies (Propositions 1 and 3) while also providing novel, comparative-statics predictions (Propositions 2 and 4). In the next two sections we report on experimental work which was conducted to test the model's predictions and assumptions (specifically the strength-of-preference assumption) in the context of the preference reversal phenomenon discussed in the previous section.

4.1 Experimental Design

We conducted five experiments with a total of 503 subjects. Table 1 provides an overview of all experiments. Each experiment consisted of two parts. The first part was used to estimate individual certainty equivalents for each subject. The second part was the actual preference reversal experiment consisting of an evaluation phase and a choice phase. The choice phase was identical across all experiments, but the estimation and evaluation phases differed between experiments as explained below.

The goal of the first part (estimation) was to obtain a measure of each subject's individual preference. Subjects faced 32 lottery pairs that were unrelated to the P-bet/$-bet pairs used in the second part of the experiment (preference reversal experiment).[2] We used each subject's elicited preferences over these 32 lottery pairs to estimate an individual utility function (see Section 4.3 below). This was done out of sample in the sense that the estimation relied exclusively on the preferences over the 32 lottery pairs from this first part, but was used as an external measure of subjects' certainty equivalents for the P-bets and $-bets used in the preference reversal experiment in the second part. In *RANK2*, *WTA2*, and *WTP*, we additionally elicited WTA or WTP valuations also for the 64 lotteries used in the first part (estimation). In Section 6.2, we use these valuations to show that our results are robust when utilities are estimated out of monetary evaluations instead of choices.

[2]See Appendix Appendix F for the complete list of lotteries used in the experiments. In the Cologne experiments the first part also included four pairs with dominated choices as a consistency check, but across both experiments subjects made only 5 dominated choices (out of $190 \times 4 = 760$). Consequently, we decided not to include them in the Zurich experiments.

The preference reversal experiments (second part) consisted of a choice phase and an evaluation phase. Each experiment elicited subjects' preferences over 60 P-bet/$-bet pairs using two different elicitation methods (choice and evaluation). All lotteries were of the form (p, x), that is, a lottery pays an amount x with probability p and zero otherwise. Lottery pairs were constructed such that the P-bet pays a moderate amount with a high probability well above 50%, whereas the $-bet pays a high amount with a much lower probability well below 50%. In the choice phase, for each of the 60 pairs subjects were asked to choose whether they would prefer to play the P-bet or the $-bet. In the evaluation phase, subjects evaluated the same 120 lotteries (60 P-bets and 60 $-bets) using a different elicitation method that differed across experiments.

In experiments *RANK1* and *RANK2*, the evaluation phase used a ranking-based elicitation procedure. Lotteries were presented in blocks of six, and subjects were asked to rank the six lotteries from their most (rank 1) to their least preferred (rank 6) option. Each block contained three P-bet/$-bet pairs. In experiments *WTA1* and *WTA2*, the evaluation phase used willingness-to-accept (WTA) valuations. Specifically, subjects were asked to state their minimal selling price for each of the 120 lotteries. Lotteries were presented sequentially on separate screens and in randomized order. Experiment *WTP* used willingness-to-pay (WTP) valuations in the evaluation phase and was otherwise identical to *WTA2*.

4.2 Procedures

Experiments *RANK1* and *WTA1* were conducted at the University of Cologne (Germany), and *RANK2*, *WTA2*, and *WTP* were conducted at the University of Zurich (Switzerland). Participants were recruited from the respective student populations, excluding students majoring in psychology or economics (who might have learned about the preference reversal phenomenon) and subjects who had previously participated in experiments involving lottery choice. The experiments in Cologne and Zurich were computerized using PsychoPy (Peirce, 2007) and z-Tree (Fischbacher, 2007), respectively.

Lotteries were presented in the form of colored pie charts, with colors (green and blue) counterbalanced across subjects. The screen position (left or right) of lotteries within pairs was also counterbalanced within subjects, with half of the pairs displaying a $-bet on the right. To control for order effects, each subject was randomly assigned to one of four different, pre-randomized sequences of lottery pairs.[3]

Before beginning the experiment, subjects were provided with general instructions and had to answer four control questions to ensure their understanding of the concept of a lottery and its pie-chart representation. Detailed instructions for all parts were presented on-screen before the start of the respective task. At the end of the experiment, subjects were asked to complete a short questionnaire eliciting various demographics (gender, age, field of studies).[4] There was no feedback during the course of the experiment, that

[3]We found no evidence for order effects on our main variables of interest.

[4]The Cologne experiments also elicited numerical literacy (Lipkus et al., 2001).

is, subjects did not receive any information regarding their earnings until the very end of the experiment. All decisions were made independently and at a subject's individual pace.

Payment procedures were explained within the instructions and carried out truthfully. To determine a subject's payoff, one lottery from each phase was randomly selected and paid (Azrieli et al., 2018). For the choice phase and the (choice-based) estimation phase the payment mechanism was identical in all experiments; one of the lottery pairs was randomly selected and the lottery chosen by the participant was played out.

For *RANK1*, *RANK2*, and *WTA1*, the evaluation phase used a variant of the (incentive-compatible) Ordinal Payment Method (OPM; Goldstein and Einhorn, 1987; Tversky et al., 1990; Cubitt et al., 2004). The computer selected one round/block at random, and then randomly selected two of the six lotteries in the round/block.[5] The one that the participant had evaluated higher was then played out.

For experiment *WTA1*, we chose the OPM because it is more intuitive than the Becker-DeGroot-Marschak (BDM) procedure (Becker et al., 1964). The latter has also been found to be noisier (Alós-Ferrer et al., 2016). A potential concern with the OPM, though, is that it only fully incentivizes subjects to truthfully reveal the ordinal ranking of their valuations but not the actual levels. To ensure that the interpretation of differences between elicited valuations and estimated certainty equivalents is justified, and as an additional robustness check, experiments *WTA2* and *WTP* relied on the BDM procedure, which incentivizes subjects to state their true valuations. For the evaluation phase, the computer selected one lottery at random. For that lottery, the computer then randomly selected an offer price between zero and the highest outcome of the lottery. The lottery was played out if the offer price was smaller than the subject's stated WTA/WTP and the subject received the offer price otherwise. For the additional evaluations in the first part (estimation) of experiments *RANK2*, *WTA2*, and *WTP*, one lottery pair was randomly selected and paid according to the BDM procedure.

The total payoff from the experiment was the sum of the amounts received from the estimation, evaluation, and choice phases. In addition subjects received a lab-mandated show-up fee of €4 in Cologne and CHF 10 in Zurich. The average total remuneration was €19.76 in Cologne and CHF 30.62 in Zurich. Sessions lasted between 70 and 85 minutes including instructions and payment.

4.3 Description of the Estimation Procedure

The estimation phase was only used to estimate subjects' individual preferences out-of-sample. The 32 lottery pairs used in this phase were constructed to maximize the precision of the estimated preferences. To achieve this we relied on optimal design

[5]To ensure comparability with *RANK1* and *RANK2*, in *WTA1* the lotteries were also presented in 20 "rounds," separated by screens announcing the next round. Each such round consisted of six lotteries presented sequentially, with the set of lotteries in a round corresponding to one block of experiments *RANK1* and *RANK2*.

theory (Silvey, 1980) in the context of non linear (binary) models (Ford et al., 1992; Atkinson, 1996), in agreement with the recommendations of Moffatt (2015).[6]

We assume that the structure of errors follows an additive random utility model (e.g., Thurstone, 1927; Luce, 1959; McFadden, 2001). However, all results throughout the paper remain qualitatively unchanged if we adopt a random preference model (Loomes and Sugden, 1998; Apesteguía and Ballester, 2018) instead (see Section 6.3). Estimation of individual risk attitudes relies on a well-established maximum likelihood procedure (e.g., see Train, 2003; Moffatt, 2005; Bellemare et al., 2008). We refer the interested reader to Appendix Appendix A for a detailed description of the estimation procedure.

For the functional form of the estimated utilities, we adopt the normalized constant absolute risk aversion (CARA) function as in Conte et al. (2011), which is given by

$$u(x) = \begin{cases} \frac{1-\exp(-rx)}{1-\exp(-rx_{\max})}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases}$$

where $x_{\max}$ is the upper bound of the outcome variable x. All our results remain qualitatively unchanged if we assume a CRRA utility function instead (see Section 6.3).

5 Results

In this section, we discuss the results of experiments *RANK1*, *RANK2*, *WTA1*, and *WTA2*. Experiment *WTP* served as a robustness check, which we discuss in Section 6.1. For all four experiments, we use the choice-based, out-of-sample estimates of subjects' individual certainty equivalents derived from the 32 binary lottery choices of the first part (estimation). Section 6.2 presents a robustness analysis, which shows that our results do not hinge on estimating certainty equivalents out of choices but instead are qualitatively unchanged when certainty equivalents are estimated from evaluations instead. Recall that in *WTA1* valuations were incentivized with an ordinal payment method, whereas in experiment *WTA2* valuations were incentivized using the Becker-DeGroot-Marschak procedure. Hence, a comparison of those two experiments also allows us to study robustness with respect to how valuations are incentivized.

Each experiment delivers three types of data for each of the 60 P-bet/$-bet pairs $(\mathrm{P}_k, \$_k)_{k=1}^{60}$ and each subject i: First, a binary choice function that takes the value 1 if P_k was chosen over $\$_k$ in the choice phase and 0 otherwise; second, a binary evaluation function that takes the value 1 if P_k was evaluated higher than $\$_k$ in the evaluation phase and 0 otherwise; and third, an out-of-sample estimate of the certainty equivalent difference between P_k and $\$_k$. In the remainder of this section, we use this data to test

[6] We chose to estimate risk attitudes from a sequence of pairwise lottery choices over alternatives such as the multiple price list (MPL) method (Holt and Laury, 2002). The reason is that the latter imposes a strong correlation structure on the choice sequence, namely a unique switching point (see Andersen et al., 2006, for a discussion of the weaknesses of MPL methods). Moreover, Beauchamp et al. (2019) show that MPL methods are susceptible to the compromise effect, which may lead to biased results.

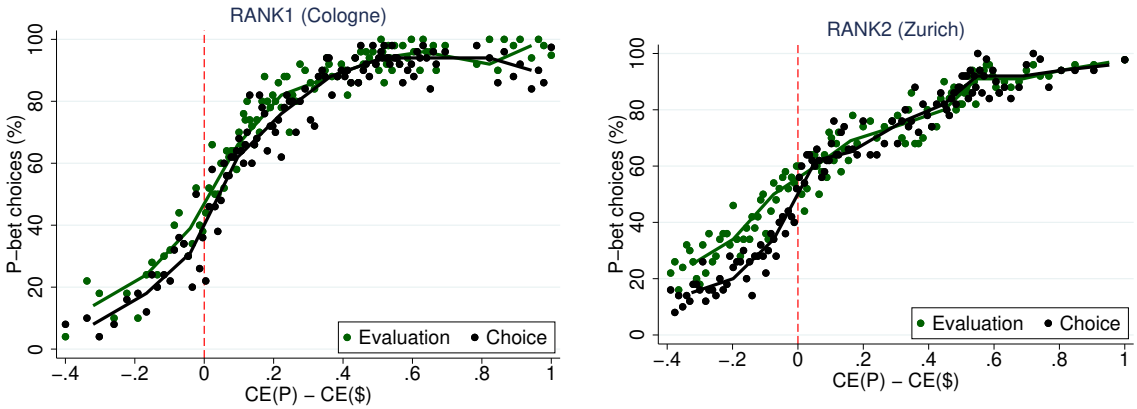


Figure 3: Strength-of-preference effects in *RANK1* (left) and *RANK2* (right).

the assumptions and the predictions of the model laid out in Section 2 at the population level.

5.1 Unbiased Evaluations (*RANK1* and *RANK2*)

For each experiment, we consider the *empirical stochastic choice function* $\hat{\rho}_c$ and the *empirical stochastic valuation function* $\hat{\rho}_v$ as a function of the estimated difference in CEs $\hat{\Delta}_i(\mathrm{P}_k, \$_k)$. That is, for a given interval $[\underline{\Delta}, \bar{\Delta}]$ those functions give the proportion of choices and evaluations, respectively, that favor the P-bet. Figure 3 plots the empirical stochastic choice and evaluation functions for *RANK1* (left) and *RANK2* (right) binning observations.[7] For both choices and evaluations, we find a monotonically increasing, sigmoidal relation with the P-bet being chosen and evaluated higher more often for larger CE differences Δ. Thus, we find clear evidence for SoP effects in choices and evaluations.

For both experiments, the empirical stochastic evaluation function exhibits no systematic shift relative to the empirical stochastic choice function. Hence, the ranking-based elicitation method used in the evaluation phase of *RANK1* and *RANK2* is an example of an unbiased evaluation method in the sense that $\hat{\rho}_c \simeq \hat{\rho}_v$. In *RANK1*, the proportion of P-choices was $\pi_c^1 = 0.68$, whereas the proportion of evaluations that ranked the P-bet above the $-bet was $\pi_v^1 = 0.72$. Analogously, we have $\pi_c^2 = 0.55$ and $\pi_v^2 = 0.59$ for *RANK2*. That is, both experiments are biased toward P-bets and, thus, Proposition 1 predicts a type-1 anomaly with more non-standard than standard reversals.

To test this prediction, we turn to the reversal rates. Overall, the rate of reversals was relatively low, amounting only to 19.46% in *RANK1* and 26.27% in *RANK2*. Figure 4 displays violin plots for individual rates of standard and non-standard reversals, for both experiments. We find that the rate of non-standard reversals in *RANK1* (*RANK2*) is

[7] To construct the bins, we first order all observations by their individual difference $\Delta_i(P_k, \$_k)$. Starting at the 'zero bin,' which contains the 50 observations closest to zero, we symmetrically form bins of up to 50 observations.

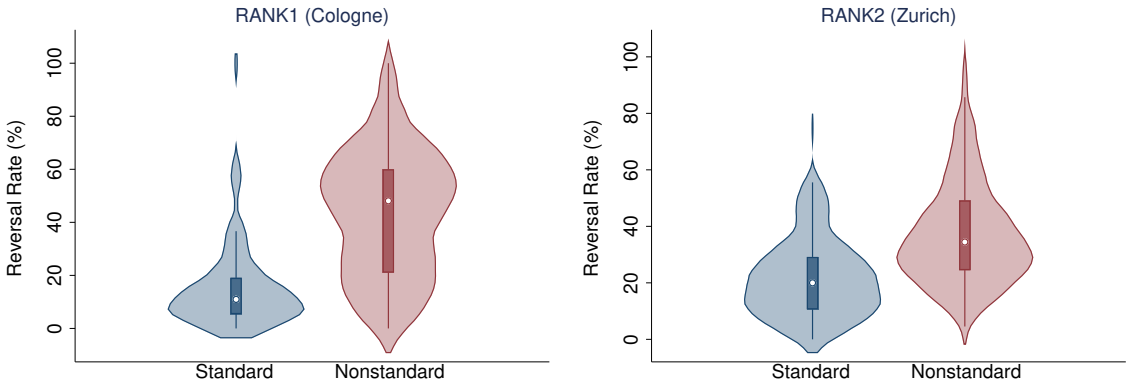


Figure 4: Rates of standard and non-standard reversals for *RANK1* (left) and *RANK2* (right). Violin plots show the median, the interquartile range and the 95% confidence intervals as well as rotated kernel density plots on each side.

44.41% (38.56%), which is higher than the rate of standard reversals of 15.73% (22.18%) confirming the prediction of Proposition 1 (WSR tests; *RANK1*: $N = 86$, $z = -5.893$, $p < 0.001$; *RANK2*: $N = 104$, $z = -5.026$, $p < 0.001$).[8]

That is, using a ranking-based elicitation method leads to the so-called "reversal of the preference reversal phenomenon" (Casey, 1991; Alós-Ferrer et al., 2016), which so far has been considered a puzzle. Our stochastic choice model provides an explanation for the occurrence of the reversal of the preference reversal phenomenon, which is confirmed by the data. Far from resulting from a behavioral bias, this phenomenon is merely a consequence of stochastic choice and the experiment being biased toward P-bets. The latter is a consequence of a combination of the particular way in which preference reversal experiments are designed and the fact that standard experimental populations are on average risk-averse. In typical experiments of this kind, lottery pairs $(\mathrm{P}_k, \$_k)$ are constructed in such a way that the expected values of P_k and $\$_k$ are similar, but $\$_k$ is riskier. Since decision makers tend to be risk averse, it follows that for the majority of lottery pairs the differences in certainty equivalents are positive. Hence, in the absence of a bias in evaluations, risk aversion leads to a bias in the experiment toward P-bets, which in turn leads to the reversal of the preference reversal phenomenon (a type-1 anomaly).

Indeed, in *RANK1* (*RANK2*) the majority of subjects is risk averse: only 13 (11) subjects or about 13.68% (10.19%) are classified as (mildly) risk-seeking. In *RANK1* (*RANK2*) the average estimated risk propensity, $\hat{r}$, is 0.152 (0.036) with a median of 0.160 (0.034) and a standard deviation of SD 0.102 (0.031).[9]

[8] Tests for differences in reversal rates can only include subjects for which both rates can be computed. For subjects with very few P-bet or $-bet choices, reversal rates tend to be on the extremes at 0% or 100%. Therefore, when calculating standard and non-standard reversal rates we only include subjects with at least four choices of each type. We obtain qualitatively the same results when all subjects for which rates can be computed are used in the analysis.

[9] An agent with a risk propensity equal to the average in *RANK1* (*RANK2*) would have a certainty equivalent of about \$3.25 (\$4.553) when facing a lottery paying \$10 with 50% probability and zero otherwise.

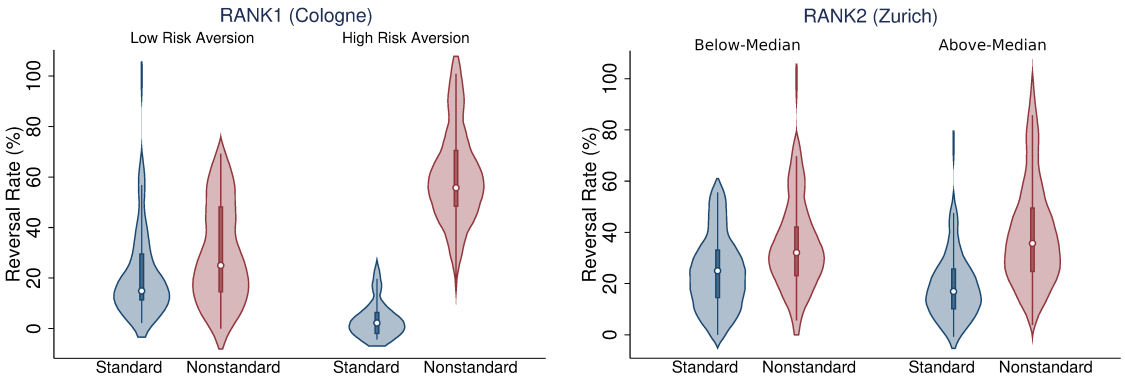


Figure 5: Reversal rates for median split on risk aversion (median split) in *RANK1* (left) and *RANK2* (right).

Proposition 2(a) provides a novel, comparative-statics prediction on how differences in individual characteristics affect the relation between the rates of standard and non-standard reversals for a given set of lottery pairs D. Specifically, the prediction is that the ratio SR/NR should be decreasing in the extent of the bias in experiment D toward P-bets (captured by $\pi_c(D)$). To test this prediction, we conducted a median split of subjects in each experiment according to their individually-estimated risk attitudes. Indeed, for the high risk aversion group in *RANK1* (*RANK2*) the proportion of P-bet choices is 80.10% (59.78%), whereas it is only 56.31% (50.22%) for the low risk aversion group. These differences are significant according to Mann-Whitney-Wilcoxon (MWW) tests (*RANK1*: $N = 95$, $z = 6.022$, $p < 0.001$; *RANK2* $N = 108$, $z = 2.626$, $p = 0.008$). That is, for the former group the bias in the experiment toward P-bets is stronger than for the latter group. Thus Proposition 2(a) predicts a smaller ratio of standard to non-standard reversals for the high risk aversion group than for the low risk aversion group. Figure 5 shows the reversal rates for the two groups for *RANK1* (left) and *RANK2* (right). In the low risk aversion group the rates of standard and non-standard reversals for *RANK1* (*RANK2*) are 22.13% (24.59%) and 29.87% (35.96%), respectively. In contrast, in the high risk aversion group the average rates of standard and non-standard reversals for *RANK1* (*RANK2*) are 8.03% (19.86%) and 61.92% (41.06%), respectively. The ratio of standard to non-standard reversals is 0.741 (0.684) for the low risk aversion group and 0.130 (0.484) for the high risk aversion group. The differences are statistically significantly (MWW tests; *RANK1*: $N = 84$, $z = -5.930$, $p < 0.001$; *RANK2*: $N = 107$, $z = -2.069$, $p = 0.038$), in line with Proposition 2(a). That is, risk aversion exacerbates type-1 anomalies in experiments that rely on unbiased evaluation methods like the ranking-based method employed in *RANK1* and *RANK2*.

Further, Proposition 2(b) predicts that the ratio SR/NR is decreasing as the bias in an experiment D toward P-bets becomes stronger ($\pi_c(D)$ increases). We can test this prediction in two ways: First, by comparing individual lottery pairs within *RANK1*

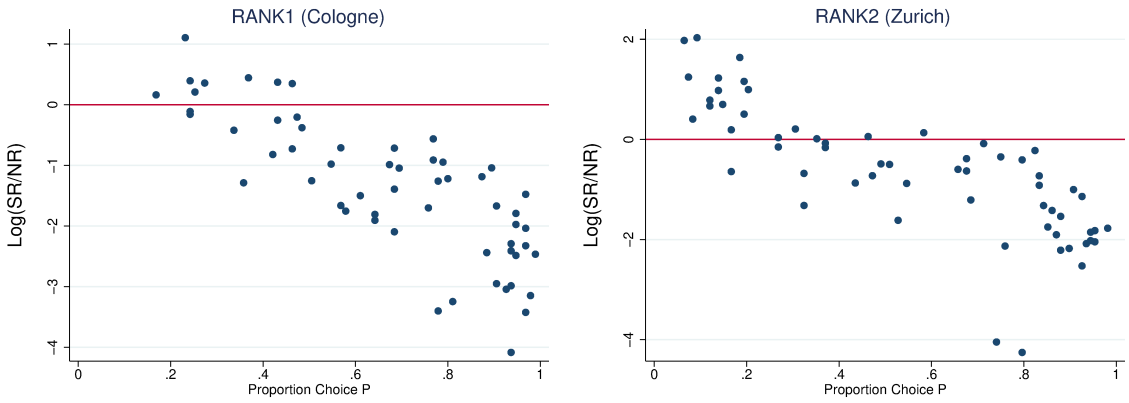


Figure 6: Average ratio between standard and non-standard reversal rates (on a log scale) by lottery pair as a function of the proportion of P-bet choices for *RANK1* (left) and *RANK2* (right).

and *RANK2*, respectively. Second, by comparing *RANK1* to *RANK2*. Figure 6 plots the average proportion of P-bet choices $\pi_c(\{\Delta_k\})$ against the average ratio SR/NR (on a log scale), separately for each lottery pair $(\mathrm{P}_k, \$_k)$. A negative correlation between the SR/NR ratio and $\pi_c(\{\Delta_k\})$ is evident (Spearman; *RANK1*, $\rho = -0.814$, $N = 58$, $p < 0.001$; *RANK2*, $N = 60$, $\rho = -0.848$, $p < 0.001$), in line with Proposition 2(b). Further, comparing across experiments, we have $\pi_c^1(D) = 0.68 > 0.55 = \pi_c^2(D)$. That is, the bias in *RANK1* toward P-bets is stronger than in *RANK2* (MWW; $N = 203$, $z = 4.760$, $p < 0.001$). In line with Proposition 2(b), we find that the SR/NR ratio is 0.314 in *RANK1* and thus smaller than the ratio of 0.633 in *RANK2*. The difference is significant (MWW; $N = 118$, $z = 3.037$, $p = 0.002$).

5.2 Biased Evaluations (*WTA1* and *WTA2*)

We now consider the two experiments that used willingness-to-accept valuations in the evaluation phase. Figure 7 plots the empirical stochastic choice and evaluation functions for *WTA1* (left) and *WTA2* (right). For choices and evaluations we again find a monotonically increasing, sigmoidal relation between the propensity to choose the P-bet and the difference in certainty equivalents in both experiments. Thus, also in *WTA1* and *WTA2* we find support for SoP effects in choices and evaluations.

Similarly to the ranking experiments, the empirical stochastic choice function is roughly symmetric around zero and takes the value one half for CE differences close to zero. However, in contrast to the two ranking-based experiments, the stochastic evaluation function is clearly shifted downwards relative to the stochastic choice function, taking a value well below one half around zero (recall Figure 2, left). Even for relatively large differences in CE, the propensity to evaluate the P-bet higher than the \$-bet barely reaches 50%. Hence, in experiments *WTA1* and *WTA2* the DMs exhibit a \$-bias in evaluations elicited via willingness-to-accept valuations. In *WTA1*, the proportion

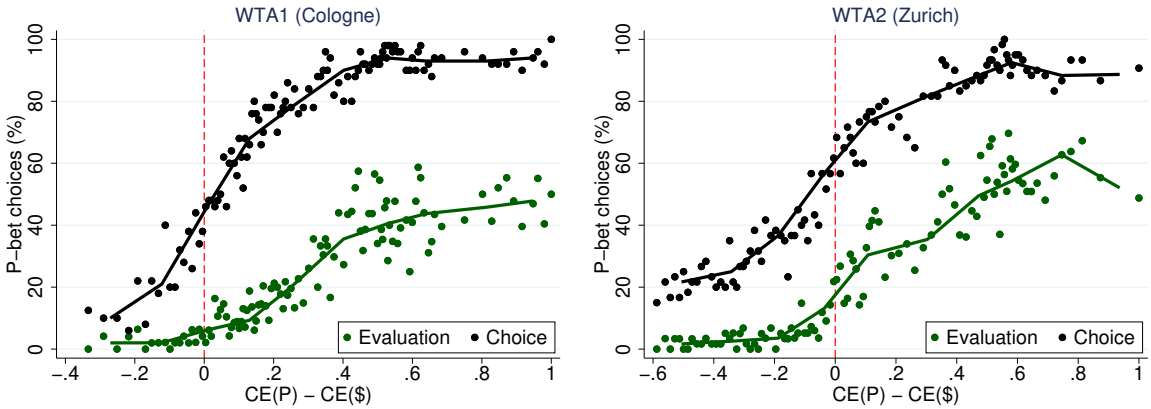


Figure 7: Strength-of-preference effects in *WTA1* (left) and *WTA2* (right).

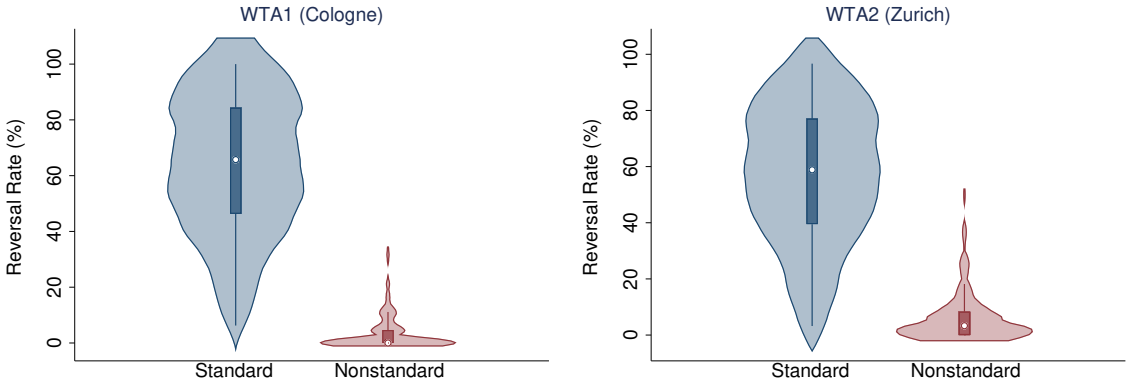


Figure 8: Reversal rates for *WTA1* (left) and *WTA2* (right).

of P-choices was $\pi_c^1(D) = 0.70$, whereas the proportion of evaluations that assigned a larger WTA to the P-bet than to the \$-bet was $\pi_v^1(D) = 0.24$. Analogously, we have $\pi_c^2(D) = 0.60$ and $\pi_v^2(D) = 0.26$ for *WTA2*. Both experiments are biased toward P-bets and $\pi_c^i(D)$ is significantly larger than $\pi_v^i(D)$ (MWW tests; *WTA1*: $N = 95$, $z = 8.441$, $p < 0.0001$; *WTA2*: $N = 103$, $z = 8.644$, $p < 0.0001$). For an unbiased experiment ($\pi_c(D) = \frac{1}{2}$), Proposition 3 predicts that if DMs exhibit a \$-bias in evaluations, then a type-2 anomaly with more standard than non-standard reversals is expected to occur. Experiments *WTA1* and *WTA2* are not unbiased, and intuitively the experimental bias could dampen type-2 anomalies. Since type-2 anomalies are ubiquitous in preference reversal experiments, though, we expect the prediction to hold if the experimental bias is not too extreme.

To test this prediction, we again turn to the reversal rates. Reversals are extremely frequent, with an average individual reversal rate of 50.63% in *WTA1* and 41.74% in *WTA2* (not distinguishing types of reversals).[10] Figure 8 displays violin plots for the

[10]Individual reversal rates were calculated excluding pairs where the P-bet and the \$-bet were identically valued. This happened in 6.74% and 5.87% of the cases in *WTA1* and *WTA2*, respectively.

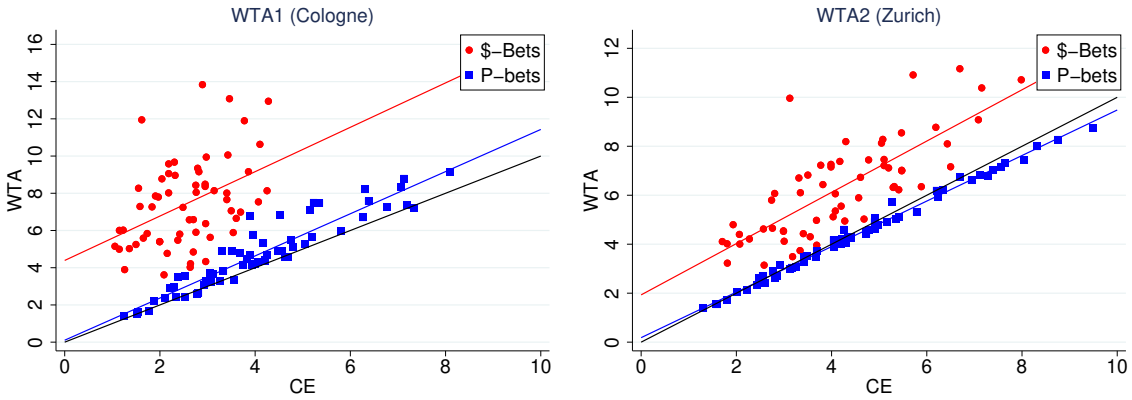


Figure 9: Accuracy of valuations in *WTA1* (left) and *WTA2* (right). Correlation between stated WTA and predicted certainty equivalent, separately for P-bets and $-bets. Each point corresponds to one lottery representing the average WTA and the average CE across all subjects for a given experiment.

individual rates of standard and non-standard reversals, for both experiments. In *WTA1* (*WTA2*) the rate of standard reversals, SR, is 63.02% (56.03%) and clearly exceeds the rate of non-standard reversals, NR, which only amounts to 3.66% (6.11%). That is, when the P-bet was chosen the propensity to state an inconsistent WTA ordering is higher than when the $-bet was chosen (WSR tests; *WTA1*: $N = 86$, $z = 8.008$, $p < 0.001$; *WTA2*: $N = 98$, $z = 8.389$, $p < 0.001$). Thus, the data shows a pronounced type-2 anomaly with more standard than non-standard reversals despite both experiments being biased toward P-bets.

Proposition 4 delivers a novel implication that serves as a causal test of the effect of a $-bias on the asymmetry in reversal rates. Specifically, it predicts that a stronger $-bias in evaluations increases the difference between standard and non-standard reversals. To test this prediction, we require a measure that allows us to compare the extent of the $-bias across DMs. Thanks to our design, we can employ the estimated individual utility functions to *quantify* the economic magnitude of the $-bias on the subject level. To that end, we consider the difference between the stated WTA valuation and the certainty equivalent derived from each subject's utility function u_i normalized by the certainty equivalent. We then take the average over all P-bets, respectively $-bets, for each i, to obtain $\beta_i(\text{P}) = \frac{1}{K}\sum_k \frac{WTA_i(\text{P}_k) - CE_i(\text{P}_k)}{CE_i(\text{P}_k)}$ and $\beta_i(\$) = \frac{1}{K}\sum_k \frac{WTA_i(\$_k) - CE_i(\$_k)}{CE_i(\$_k)}$ as a measure of subject i's accuracy in evaluations, separately for each type of lottery.

To illustrate the accuracy of subjects' WTA valuations Figure 9 plots the stated WTAs against the estimated CEs for each of the 120 lotteries, distinguishing P-bets and $-bets. For P-bets, the correlation coefficient is close to unity (Spearman; *WTA1*: $\rho = 0.930$, $N = 60$, $p < 0.001$; *WTA2*: $\rho = 0.992$, $N = 60$, $p < 0.001$) and the (WTA,CE) pairs are tightly clustered around the regression line, which is itself close to the diagonal.

Our results are unchanged when pairs with identical valuations are included and classified as either non-reversals or reversals.

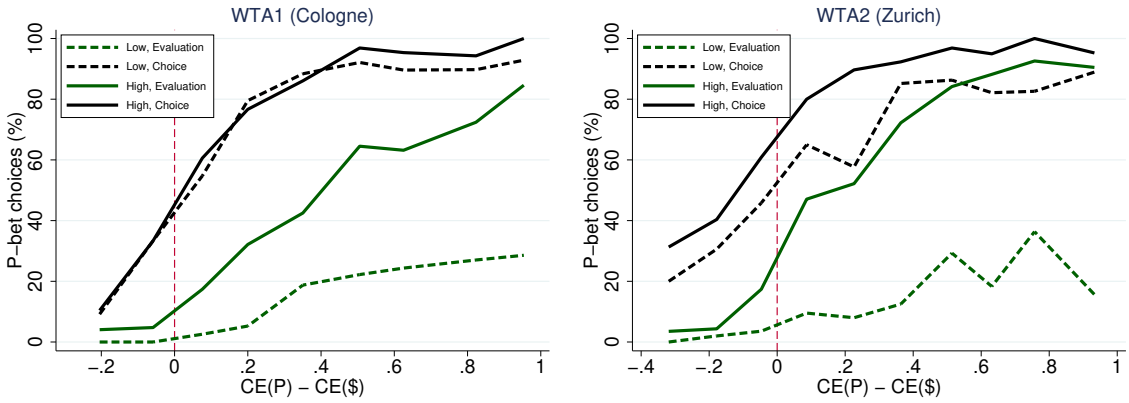


Figure 10: Median split of β_i (high versus low) for *WTA1* (left) and *WTA2* (right).

However, the same cannot be said for \$-bets, for which the picture is much more dispersed and far away from the diagonal, and the correlation is much lower (Spearman; *WTA1*: $\rho = 0.424$, $N = 60$, $p = 0.001$; *WTA2*: $\rho = 0.7644$, $N = 60$, $p < 0.001$). Thus, contrary to the generalized impression in the literature, the evaluations through stated WTA faithfully reflect certainty equivalents for P-bets derived from independently-estimated expected utilities. The implications are twofold: First, the accuracy of subjects' stated WTAs further confirms the validity of our estimated utilities. Second, there is no general, systematic difference across elicitation tasks (monetary valuations and choices) for P-bets. Crucially, we can interpret the difference $\beta_i = \beta_i(\$) - \beta_i(\text{P})$ as a direct measure of the economic magnitude of a DM's \$-bias in evaluations in monetary terms. According to this measure, we quantify the \$-bias in *WTA1* at a whopping 275% relative to the CE, and at 63% in *WTA2*.

To test Proposition 4, we now divide subjects into two groups based on a median split of their \$-bias, quantified by β_i, for each experiment. Figure 10 shows the empirical (average) stochastic choice and evaluation functions separately for the high and low \$-bias groups. For the former group, the stochastic evaluation function is shifted downwards compared to the latter, that is, the group with high values of β_i indeed exhibits a stronger \$-bias in evaluations (as defined in Section 2.2). In contrast, the stochastic choice functions are indistinguishable for *WTA1* and very close together for *WTA2*. Comparing the reversal rates for both groups, we find that a stronger \$-bias exacerbates the asymmetry between standard and non-standard reversals. Specifically, in *WTA1* (*WTA2*) the difference between SR and NR amounts to 44.6 (28.5) percentage points for the low \$-bias group, whereas it is 75.5 (69.7) percentage points for the high \$-bias group. The differences between the low and high \$-bias groups are statistically significant (MWW tests, *WTA1*: $N = 86$, $z = -4.986$, $p < 0.001$; *WTA2*: $N = 98$, $z = -6.628$, $p < 0.001$). Summarizing, we find that the asymmetry between the rates of standard and non-standard reversals is larger for subjects that exhibit a stronger \$-bias, confirming the prediction of Proposition 4.

5.3 The Classical Preference Reversal Phenomenon Revisited

Our theoretical and empirical results provide a new view on the classical preference reversal phenomenon. Ever since this phenomenon was first discovered (Slovic and Lichtenstein, 1968; Lichtenstein and Slovic, 1971; Lindman, 1971; Grether and Plott, 1979), dozens of contributions have reported the effect to be both robust and large, with considerably-higher rates of standard reversals compared to the rates of non-standard ones. Ironically, it seems that this wide consensus has hidden a misunderstanding: the effect has actually been *underestimated.* The reason is that the literature has implicitly assumed that the "default" situation in the absence of whatever causal determinants were behind the phenomenon should have been an equality in reversal rates. This is not correct. To see this, recall equation (5). For a DM exhibiting a \$-bias in evaluations we have $\pi_v(D) < \pi_c(D)$, as documented in *WTA1* and *WTA2* above. It then follows from equation (5) that $\frac{SR}{NR} > \frac{1-\pi_c(D)}{\pi_c(D)}$. In contrast, for an unbiased DM we have $\pi_v(\Delta) = \pi_c(\Delta)$, as documented in *RANK1* and *RANK2* above, which implies $\frac{SR}{NR} = \frac{1-\pi_c(D)}{\pi_c(D)}$. Hence, depending on $\pi_c(D)$, type-1 or type-2 anomalies might obtain *in the absence of any underlying behavioral bias.* The quantity $\pi_c(D)$ depends on the risk attitudes in the sample and the choice of lottery pairs in the experiment. Thus, whether the difference of standard and non-standard reversals is larger than zero or not (or whether the ratio is larger than one or not) is simply not diagnostic of behavioral biases. Simply put, a hypothetical situation with $SR = NR$ is not the proper null to study behavioral biases. Rather, this proper null is given by the equation $\frac{SR}{NR} = \frac{1-\pi_c(D)}{\pi_c(D)}$. Since most experiments are biased toward P-bets due to risk aversion and the choice of lottery pairs, we typically have that $\pi_c(D) > 1/2$ and the true default situation is one where the rates of non-standard reversals are *higher* ($SR < NR$), and hence the fact that they become lower ($SR > NR$) in preference-reversal experiments with monetary valuations shows that the underlying determinants are stronger than implicitly assumed.

We can illustrate this insight using the individual-level data from our experiments. Figure 11 plots the individual reversal ratios against the individual choice odds (both on a log scale) for the subjects in the WTA-based experiments (left) and the ranking-based experiments (right). For all but one subject in *WTA1* (seven in *WTA2*) the reversal ratio is larger than the corresponding choice odds, in line with DMs exhibiting a \$-bias in WTA-valuations. In contrast, for *RANK1* and *RANK2* the points are clustered around the diagonal, that is, for most DMs the reversal ratio closely matches the corresponding choice odds, in line with DMs being unbiased. The actual preference reversal phenomenon is the fact that the data in the WTA experiments is shifted upwards with respect to the diagonal line, not with respect to a zero level. Note that, if one would insist on evaluating possible behavioral biases with respect to a zero level, the right-hand-side of Figure 11 would suggest a P-bias for a large part of the subjects in RANK experiments (and a \$-bias for the rest), while in reality there are no behavioral biases in this case.

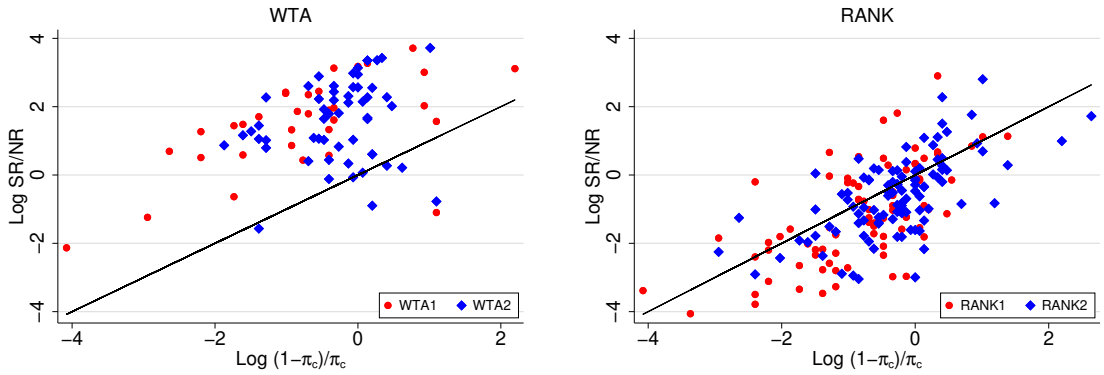


Figure 11: Figure plots individual reversal ratio $\frac{SR}{NR}$ (on a log scale) against the individual choice odds $\frac{1-\pi_c}{\pi_c}$. Left: *WTA1* and *WTA2*. Right: *RANK1* and *RANK2*.

5.4 Reexamining Previous Experiments

As outlined in the previous subsection, the proper null for a preference reversal experiment is not an equality in reversal rates. Instead, we can interpret the discrepancy between the reversal ratio $\frac{SR}{NR}$ and the choice odds $\frac{1-\pi_c(D)}{\pi_c(D)}$ as an indicator of a \$-bias in evaluations. Using this insight, we now proceed to reexamine existing preference reversal experiments from the literature (summarized in Table 2), following an analogous approach at the experiment level instead of at the subject level. Specifically, we reconsider 28 previous preference reversal experiments $\{D_1, \ldots, D_{28}\}$ and compute the average (across subjects) of $\pi_c(D_j)$, $SR(D_j)$, and $NR(D_j)$ for each experiment D_j. Figure 12 plots the reversal ratio $\frac{SR(D_j)}{NR(D_j)}$ and the choice odds $\frac{1-\pi_c(D_j)}{\pi_c(D_j)}$ for $j = 1, \ldots, 28$ on a log scale. Additionally, the figure includes our five experiments *WTA1*, *WTA2*, *RANK1*, *RANK2*, and *WTP*. Points above the horizontal dashed line are those which reported the preference reversal phenomenon, whereas points below show the reversal of the phenomenon. Points left of the vertical dashed line indicate experiments that are biased toward P-bets ($\pi_c(D_j) > \frac{1}{2}$), whereas points to the right indicate experiments that are biased toward \$-bets. The majority of the experiments (22 out of 28) exhibit the classic asymmetry with more standard than non-standard reversals, two experiments find basically identical rates of reversals of both types, and four experiments find a reversal of the preference reversal phenomenon. The majority of the experiments (16 out of 28) are biased toward P-bets, one experiment is unbiased, and eleven experiments are biased toward \$-bets. As explained above, however, only whether a point is above or below the *diagonal* is indicative of a behavioral \$-bias. Comparing choice odds and reversal ratios across experiments allows us to obtain further insights on the extent of the bias for different evaluation methods.

Experiments that elicited evaluations using WTA-valuations (shown as red points) consistently find evidence for the preference reversal phenomenon. Interestingly, however, almost all of those are above the diagonal, in line with DMs exhibiting a \$-bias in WTA valuations. The only exception is an experiment in Lichtenstein and Slovic (1973,

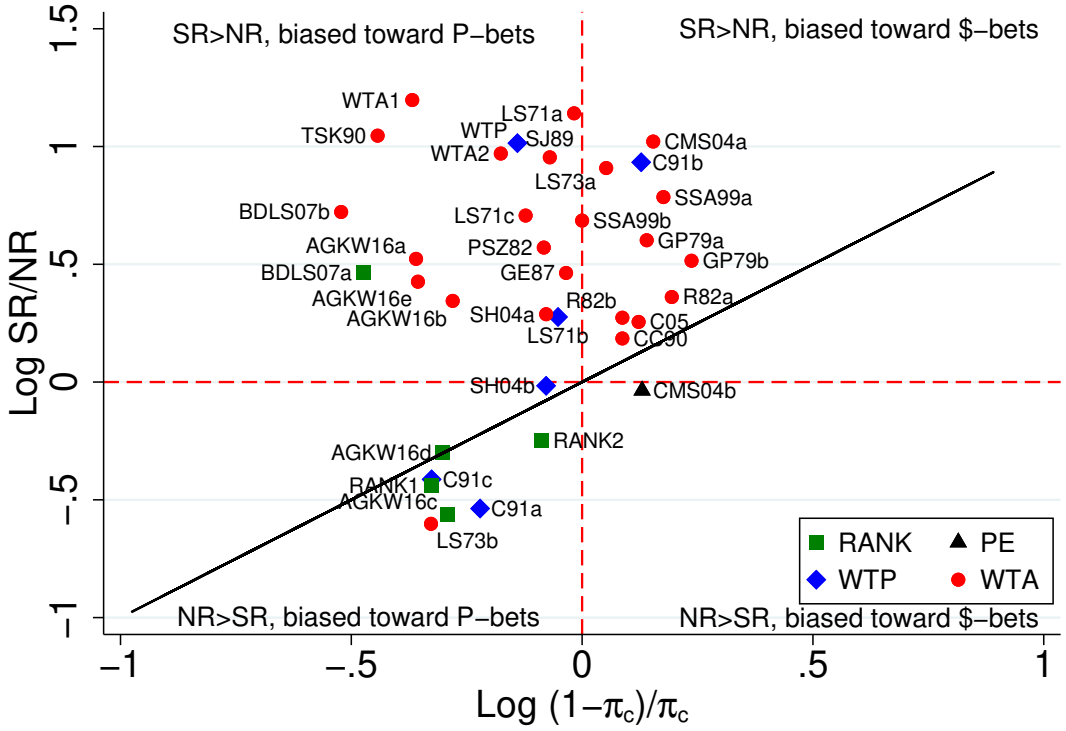


Figure 12: Logarithm of the aggregate reversal ratio $\frac{SR}{NR}$ against logarithm of the aggregate choice odds $\frac{1-\pi_c}{\pi_c}$ for different 28 previous experiments in the literature (see Table 2) and our five new experiments.

LS73b), which used only lotteries with negative expected value that involved both gains and losses.

Figure 12 also displays five experiments that used WTP-valuations (Willingness To Pay) instead of WTA-valuations (see Table 2) and came to mixed conclusions. For this reason, we carried out our own experiment using WTP, and we discuss this valuation method in Section 6.1 below.

A few experiments have used rank-based evaluation tasks analogous to *RANK1* and *RANK2* instead of monetary elicitations. Bateman et al. (2007) indirectly inferred monetary valuations using a task that asked subjects to rank P-bets and $-bets separately but together with sure amounts (BDLS07a). They found a smaller asymmetry in reversal rates compared to a second experiment using WTA-valuations (BDLS07b). Both experiments were biased toward P-bets to a similar extent (π_c(BDSL07b) = 0.75, π_c(BDSL07a) = 0.77). Thus, our results suggest that DMs exhibited a $-bias also in their ranking-based task but that it was weaker than for WTA-valuations. Going one step further, Alós-Ferrer et al. (2016) used a pure ranking task (AGKW16c,d), which led to the reversal of the preference reversal phenomenon, whereas the preference reversal phenomenon was observed in other treatments that used WTA-valuations (AGKW16a,b,e).

Notably, (AGKW16d) sits squarely on the diagonal, in full agreement with our conclusions that with this method DMs are unbiased.

Our analysis also suggest a new interpretation of some previous results. For example, in light of the scale compatibility hypothesis, it is natural to speculate that one could induce a P-bias, as opposed to the standard \$-bias observed for WTA-valuations, by shifting the focus from outcomes to probabilities. For this reason, Cubitt et al. (2004) compared an experiment with WTA-valuations (CMS04a) to another (CMS04b) where valuations were made in terms of probability equivalents (PE). Both experiments were slightly biased toward \$-bets (π_c(CMS04a) = 0.41 and π_c(CMS04b) = 0.43). While WTA-valuations led to the preference reversal phenomenon, however (and disappointingly), there was no difference between reversal rates for PE-valuations. The fact that valuations through probability equivalents did not "flip" the phenomenon thus cast doubts on the compatibility hypothesis. However, our analysis suggests that the results of Cubitt et al. (2004) could be interpreted exactly as originally intended. The key is that, again, the absence of a difference between reversal rates is *not* diagnostic unless the experiment is unbiased ($\pi_c = \frac{1}{2}$). Figure 12 shows that experiment CMS04b is actually located below the diagonal, which in view of our results suggests that indeed subjects exhibited a (small) P-bias for the PE method.

In summary, our framework accommodates previous findings from the literature and provides explanations for several observations which were so far considered inconsistent or hard to explain. This is made possible by taking into account the combination of differences in the bias in evaluations and differences in the extent to which specific experiments are biased toward P-bets or \$-bets.

Table 2: Overview of preference reversal experiments in the literature.

Study	Label	Evaluation	π_c	SR	NR
Lichtenstein and Slovic (1971), Exp1	LS71a	WTA	0.51	0.83	0.06
Lichtenstein and Slovic (1971), Exp2	LS71b	WTP	0.53	0.51	0.27
Lichtenstein and Slovic (1971), Exp3	LS71c	WTA	0.57	0.56	0.11
Lichtenstein and Slovic (1973), PosEV	LS73a	WTA	0.47	0.81	0.10
Lichtenstein and Slovic (1973), NegEV	LS73b	WTA	0.68	0.19	0.76
Grether and Plott (1979), Exp1	GP79a	WTA	0.42	0.56	0.14
Grether and Plott (1979), Exp2	GP79b	WTA	0.37	0.68	0.21
Pommerehne et al. (1982)	PSZ82	WTA	0.55	0.47	0.13
Reilly (1982), Exp1	R82a	WTA	0.39	0.62	0.27
Reilly (1982), Exp2	R82b	WTA	0.45	0.30	0.16
Goldstein and Einhorn (1987)	GE87	WTA	0.52	0.61	0.21
Schkade and Johnson (1989)	SJ89	WTA	0.54	0.54	0.06
Chu and Chu (1990)	CC90	WTA	0.45	0.49	0.32
Tversky et al. (1990)	TSK90	WTA	0.74	0.45	0.04
Casey (1991), Exp1	C91a	WTP	0.63	0.21	0.71
Casey (1991), Exp2	C91b	WTP	0.43	0.85	0.10
Casey (1991), Exp3	C91c	WTP	0.68	0.20	0.53
Selten et al. (1999), Exp1	SSA99a	WTA	0.40	0.61	0.10
Selten et al. (1999), Exp2	SSA99b	WTA	0.50	0.63	0.13
Cubitt et al. (2004), Exp1	CMS04a	WTA	0.41	0.35	0.03
Cubitt et al. (2004), Exp2	CMS04b	PE	0.43	0.20	0.21
Schmidt and Hey (2004), Exp1	SH04a	WTA	0.55	0.33	0.17
Schmidt and Hey (2004), Exp2	SH04b	WTP	0.55	0.26	0.27
Chai (2005)	C05	WTA	0.43	0.27	0.15
Bateman et al. (2007), Exp1	BDLS07a	WTA	0.77	0.26	0.05
Bateman et al. (2007), Exp2	BDLS07b	RANK	0.75	0.17	0.06
Alós-Ferrer et al. (2016), Exp1, BDM	AGKW16a	WTA	0.70	0.50	0.15
Alós-Ferrer et al. (2016), Exp1, OPM	AGKW16b	WTA	0.66	0.42	0.19
Alós-Ferrer et al. (2016), Exp2, Unframed	AGKW16c	RANK	0.66	0.18	0.55
Alós-Ferrer et al. (2016), Exp2, Framed	AGKW16d	RANK	0.67	0.20	0.40
Alós-Ferrer et al. (2016), Exp2, BDM	AGKW16e	WTA	0.69	0.48	0.18

6 Robustness Analyses

This section briefly reports three robustness analyses. Section 6.1 shows that we obtain qualitatively the same results if WTP valuations are used in the evaluation phase instead of WTA valuations. Section 6.2 shows that our results remain robust when alternative utility functions are estimated out-of-sample from unrelated WTA valuations. Finally, Section 6.3 shows that our results do not hinge on the specific assumptions for the estimation of utilities.

6.1 WTA versus WTP Valuations

In the main analysis, we have relied on WTA-valuations because this is the most common choice in the literature. Preference reversals have also been shown to obtain if one uses willingness-to-pay (WTP) valuations or sequential elicitation methods instead (Butler and Loomes, 2007). There are, however, some differences. For instance, it has been previously argued that WTPs might be less biased than WTAs (Schmidt and Hey, 2004). However, empirical results using WTP have found both the preference reversal phenomenon and its reversal (see Figure 12). The second experiment of Schmidt and Hey (2004, SH04b) used WTP instead of WTA (as in SH04a) in the evaluation phase. Both of their experiments used the same lotteries and were essentially unbiased ($\pi_c = 0.54$). Hence, the choice odds ratio is close to one and in the absence of a \$-bias in evaluations no asymmetry between reversal rates should be expected. While for WTA-valuations (SH04a) the standard asymmetry was observed, for WTP-valuations (SH04b) reversal rates were indistinguishable. Evaluations through WTP were also used in a series of (non-incentivized) experiments by Casey (1991). Experiments 1 and 3 (C91a, C91c) used large, hypothetical stakes and were biased toward P-bets (π_c(C91a) = 0.68 and π_c(C91c) = 0.62). Here, indeed the reversal of the preference reversal phenomenon ($NR > SR$) was observed. In contrast, Experiment 2 (C91b) used small stakes and was biased toward \$-bets (π_c(C91b) = 0.43). Here, the classical preference reversal phenomenon ($SR > NR$) was observed. Thus, evidence from these previous experiments suggests that with WTP-valuations DMs might exhibit a smaller \$-bias in evaluations than with WTA-valuations or even no bias at all. If that would be the case, then equation (5) implies that whether the preference reversal phenomenon or its reversal is observed should mainly depend on the degree to which the experiment is biased toward P-bets or \$-bets, which would explain the seemingly inconsistent findings in Casey (1991).

However, apart from using WTP valuations, the experiments by Schmidt and Hey (2004) and Casey (1991) also differed from typical preference reversal experiments in a number of other dimensions (e.g. hypothetical payoffs, number of repetitions, and number of outcomes of the lotteries). Hence, to clarify the role of WTP valuations and to verify whether they are indeed unbiased, we conducted an additional experiment, *WTP*, that was identical to *WTA2* with the sole exception that the evaluation phase relied on WTP instead of WTA. The results of experiment *WTP*, which are reported in

Appendix Appendix B, are very similar to the results of experiments *WTA1* and *WTA2* reported in Section 5. In particular, we find a \$-bias also in WTP valuations, which is of a similar magnitude (actually, slightly larger) than in *WTA2* (MWW test on β_i; $N = 207$, $z = 1.843$, $p = 0.065$). Thus, our results suggest that WTP valuations are also biased.

6.2 Price-based versus Choice-based Utility Estimation

So far we have relied on choice-based, out-of-sample estimates of subject's individual utility functions. Although we are convinced that this approach is appropriate, we acknowledge that one might argue that preferences estimated from choices (although out-of-sample) are likely to reflect other binary choices (like those in the choice phase) better than monetary valuations (like those in the evaluation phase of the WTA experiments) simply because the decision situations are more similar. In this section, we show that this is *not* the case. Intuitively, this argument suggests that one should obtain the opposite result when utilities are estimated from monetary valuations instead. Although this is an intuitive line of thought, we can clearly refute this conjecture.

In experiments *RANK2* and *WTA2*, we elicited WTA-valuations also for the 64 lotteries used for utility estimation in the first part. We then repeated the estimation exercise described in Section 4.3 using the imputed choices derived from those WTA-valuations. That is, for a lottery pair (A, B) used in the first part we consider A to be "chosen" by subject i if and only if $WTA_i(A) > WTA_i(B)$.[11] We then estimated new utility functions u_i' using these imputed, valuation-based "choices." The robustness analysis, which we report in Appendix Appendix C, replicates all our previous findings. That is, our results do not hinge on estimating utilities out of choices but instead are qualitatively unchanged when utilities are estimated from WTA-valuations instead. The reason is simply that there is no systematic bias between choices and monetary valuations in general, but rather a bias in the valuation of \$-bets only (recall Figure 9).

It is important to note, however, that in *RANK2* the empirical stochastic choice and evaluation functions based on u_i' both show a strong upwards shift, thus wrongly suggesting a P-bias in evaluations *and* choices. Further, in *WTA2* the empirical stochastic choice function is also shifted upwards, whereas the stochastic evaluation function is still shifted downwards. That is, although qualitatively the results go in the same direction, these alternative, valuation-based estimates perform poorly at accurately capturing choices or evaluations.

6.3 Alternative Utility Estimations

The results reported above relied on utilities estimated assuming a CARA functional form. As a robustness check, we repeated the RUM-based estimation exercise described

[11] In 8.44% and 7.06% of all cases in *WTA2* and *RANK2*, respectively, the stated valuation was the same for both lotteries. These observations were not considered for the estimation.

in Section 4.3 using a constant relative risk aversion (CRRA) utility function instead. We then repeated the analysis reported in Section 5 based on those CRRA estimates. The results, which we report in Appendix Appendix D, confirm that our previous observations do not hinge on the CARA specification of the utility function, but remain robust when a CRRA specification is used instead.

Random Preference Models (RPM) (e.g., Loomes and Sugden, 1995), where utility noise is replaced with parameter noise for a given functional family of utilities, have been recently defended as an alternative to standard random utility models (see Wilcox, 2008, 2011; Bruner, 2017; Apesteguía and Ballester, 2018; Vieider, 2018). As a further robustness check, we also estimated an RPM and repeated the analysis reported in Section 5 with the corresponding estimates. This analysis, which we report in Appendix Appendix E, shows that our results are also qualitatively unchanged in this case.

7 Conclusion

We have provided a stochastic choice model for decisions under risk that predicts when inconsistencies across different preference elicitation methods should occur, which kind of anomalies should be expected and when, and what determines their magnitude. This is important because inconsistencies across elicitation methods are pervasive and contradict all preference-based theories of decisions under risk. Our results show that some anomalies are not diagnostic of behavioral biases at all, but rather that they arise naturally as a consequence of regularities in stochastic choice, risk attitudes, and experimental design. Other anomalies, however, can be traced back to behavioral biases affecting the evaluation of certain alternatives. We also provide comparative-statics results showing how the anomalies are exacerbated or dampened by the strength of behavioral biases, but also by individual characteristics as risk aversion or the particularities of the alternatives used in an experiment.

To test the model, we conducted five different experiments focused on the classical preference reversal phenomenon, which is one of the most robust empirical anomalies contradicting basic microeconomic principles. The results confirm all our predictions, which include when the phenomenon occurs, when the opposite anomaly occurs instead, and how their magnitudes change. The experiments rely on the estimation of certainty equivalents using out-of-sample data, which allow us to operationalize our tests and also validate our assumptions on stochastic choice. The results are robust to the functional form used to estimate utilities, to the use of willingness-to-accept or willingness-to-pay as valuation methods, to the method used to incentivize valuations, and to the use of choices or valuations to estimate utilities.

When applied to the classical preference reversal phenomenon, our model allows us to uncover new regularities and organize the previous literature. In particular, we find that the previous literature has compared empirical results to the incorrect default, because in the absence of a behavioral bias, an anomaly is predicted which corresponds

to the reversal of the classical phenomenon. As a consequence, the magnitude of the phenomenon has actually been underestimated.

Our model and data provide a consistent, unified, systematic account of anomalies in preference elicitation including the preference reversal phenomenon. The fact that we identify the correct defaults in the absence of behavioral biases is particularly relevant for future research whenever different elicitation methods are employed. A researcher who fails to account for the correct defaults that we identify might incorrectly conclude that a behavioral bias must be invoked to explain an apparent anomaly which in reality merely reflects the correct default.

Appendix A Description of RUM Estimation

To estimate individual utility functions from the binary lottery choices in part one of the experiment we follow the approach described in Moffatt (2015, Chapter 13). All $T = 32$ trials used for the utility estimation involved binary choices between lotteries of the form $A = (p, x)$ and $B = (q, y)$, where A pays x with probability p and B pays y with probability q, and 0 otherwise. We index the trials in the experiment by $t = 1, ..., 32$, that is, in trial t subjects face the choice between $A_t = (p_t, x_t)$ and $B_t = (q_t, y_t)$. Further, we index the N subjects by $i = 1, ..., N$. In the main analysis we assume a normalized constant absolute risk aversion (CARA) function as in Conte et al. (2011), which is given by

$$u(x \mid r) = \begin{cases} \frac{1-e^{-rx}}{1-e^{-rx_{\max}}}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases} \tag{6}$$

where $x_{\max} = \max\{x_1, \ldots, x_T, y_1, \ldots, y_T\}$ is the maximum outcome across all T lottery pairs (trials). The normalization ensures that $u(x \mid r)$ is increasing also for negative values of r (indicating risk-seeking).[12] Under the assumption of Expected Utility maximization, subject i with utility function $u(x \mid r_i)$ chooses A_t over B_t if the difference in expected utilities is positive, that is,

$$\nabla_t(r_i) := p_t u(x_t \mid r_i) - q_t u(y_t \mid r_i) = \frac{p_t(1 - e^{-r_i x_t}) - q_t(1 - e^{-r_i y_t})}{1 - e^{-r_i x_{\max}}} > 0. \tag{7}$$

In order to be able to estimate the parameters of the model, we now add noise to the model. There are two standard approaches in the literature: The Fechner or Random Utility Model (RUM) and the Random Preference Model (RPM). RUM assumes that each subject is characterized by a risk parameter r_i that is fixed across trials, whereas RPM assumes that a subject's risk parameter varies randomly between trials but is drawn from a certain distribution. Since our goal is to compare certainty equivalents across multiple trials, the main analysis reported in the paper uses a RUM-based estimation.[13]

Following the RUM approach, we add an error term $\varepsilon_{it} \sim N(0, \sigma^2)$ with $\sigma^2 > 0$ to (7). That is, the lottery A_t is chosen if the following condition holds:

$$\nabla_t(r_i) + \varepsilon_{it} > 0 \tag{8}$$

Define the binary choice indicator for trial t

$$\gamma_{it} = \begin{cases} 1 & \text{if } A_t \text{ chosen by subject } i \\ -1 & \text{if } B_t \text{ chosen by subject } i. \end{cases}$$

[12] The results are qualitatively unchanged when we assume an utility function with constant relative risk aversion (CRRA) instead (see Appendix Appendix D).

[13] In Appendix Appendix E we carry out an analogue estimation using the RPM approach and report the corresponding results. We find that all results are qualitatively unchanged.

Then the probability of a choice conditional on the risk-parameter r_i is given by

$$P(\gamma_{it} \mid r_i) = P\left(\gamma_{it}\nabla_t(r_i) > \gamma_{it}(-\varepsilon_{it})\right) = P\left(\gamma_{it}\frac{\nabla_t(r_i)}{\sigma} > \gamma_{it}\frac{-\varepsilon_{it}}{\sigma}\right) = \Phi\left(\gamma_{it}\frac{\nabla_t(r_i)}{\sigma}\right) \tag{9}$$

where Φ is the standard normal cumulative distribution function.

The conditional probabilities above were derived conditional on a subject's risk parameter r_i. In other words, estimating this model over the entire population would imply homogeneity in risk attitude across subjects. In order to allow for between-subject heterogeneity, we let the risk attitudes vary across the population. In particular, we assume that the individual risk attitudes in the population are distributed normally in our subject pool according to

$$r \sim N(\mu, \eta^2).$$

Hence, the log-likelihood of a sample given by the matrix $\Gamma = (\gamma_{it})$ consisting of T trials and N subjects is

$$\log L = \sum_{i=1}^{N} \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\left(\gamma_{it}\frac{\nabla_t(r)}{\sigma}\right) f(r \mid \mu, \eta) dr \tag{10}$$

where $f(r \mid \mu, \eta) = \frac{1}{\sqrt{2\rho\eta^2}} e^{-\frac{1}{2}\left(\frac{r-\mu}{\eta}\right)^2}$ is the density function of the risk parameter r.

In order to evaluate the integral in (10) we use the method of maximum simulated likelihood (MSL) (see Train, 2003, for details). Specifically, we will approximate this integral by the following average

$$\frac{1}{H}\sum_{h=1}^{H}\left(\prod_{t=1}^{T}\Phi\left(\gamma_{it}\frac{\nabla_t(r_{ih})}{\sigma}\right)\right) \tag{11}$$

using a sequence of H (transformed) Halton draws $(r_{i1}, \ldots, r_{iH})$ from $N(\mu, \eta^2)$ for each subject i (fixed over trials t). For the estimation, we use the Stata implementation "mdraws" of this procedure (Cappellari and Jenkins, 2003). Halton draws, a by-now-standard procedure, simulate random draws that ensure even coverage of the parameter space (e.g. avoiding clustering) using Halton sequences (Halton, 1960; Moffatt, 2015). Specifically, a Halton sequence is defined for a given prime number p, for example $p = 2$, is $(\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \ldots)$. Such a sequence $(h_1, h_2, \ldots)$ provide pseudo-random draws from the uniform distribution $U(0, 1)$. To obtain draws from $N(\mu, \eta^2)$ we apply the following transformation $r_{ij} = \mu + \eta\Phi^{-1}(h_j)$ where Φ^{-1} is the inverse of the normal cumulative distribution function.

The MSL approach amounts to replacing the integral in (10) by (11) and then maximize the resulting function

$$\log \hat{L} = \sum_{i=1}^{N} \ln \frac{1}{H}\sum_{h=1}^{H}\left(\prod_{t=1}^{T}\Phi\left(\gamma_{it}\frac{\nabla_t(r_{ih})}{\sigma}\right)\right). \tag{12}$$

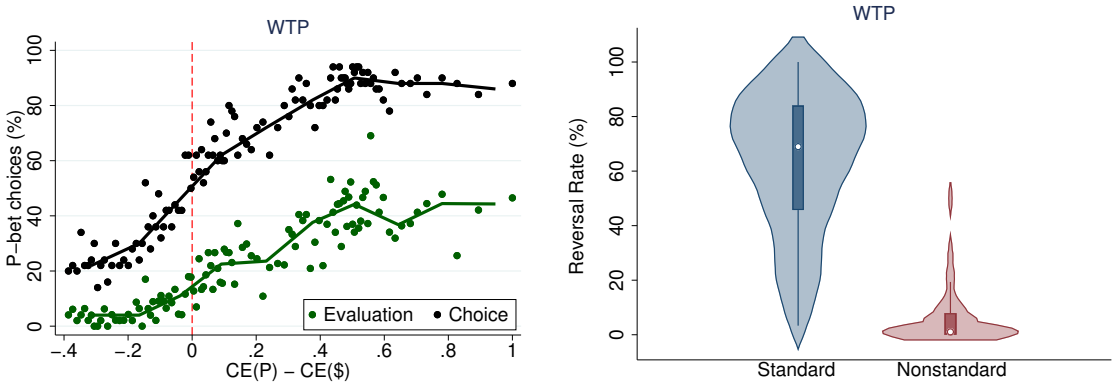


Figure A.1: Strength-of-preference effects (left) and reversal rates (right) for *WTP*.

Maximization of (12) is carried out using standard MLE routines in Stata to obtain the estimates $(\hat{\mu}, \hat{\eta}, \hat{\sigma})$. Given those estimates we obtain the posterior expectation of each subject's risk attitude $\hat{r}_i$ conditional on their T choices applying Bayes' rule as follows

$$\hat{r}_i = E(r_i|\gamma_{i1}, ..., \gamma_{iT}) \approx \frac{\frac{1}{H}\sum_{h=1}^{H} r_{ih} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{\nabla_t(r_{ih})}{\hat{\sigma}}\right)\right)}{\frac{1}{H}\sum_{h=1}^{H} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{\nabla_t(r_{ih})}{\hat{\sigma}}\right)\right)}$$

for a sequence of Halton draws $(r_{i1}, \ldots, r_{iH})$ from $N(\hat{\mu}, \hat{\eta}^2)$.

Given the estimated individual mean risk parameter $\hat{r}_i$, we obtain

$$\hat{u}_i(x) = \frac{1 - e^{-\hat{r}_i x}}{1 - e^{-\hat{r}_i x_{\max}}} \text{ for } \hat{r}_i \neq 0$$

as the estimated utility function of subject i.

Appendix B Willingness-To-Pay Valuations

Experiment *WTP* was identical to *WTA2* except that it employed willingness-to-pay (WTP) valuations instead of WTA in the evaluation phase as well as for the estimation in the first part. The experiment involved $N = 102$ subjects and it was conducted at the University of Zurich.

The left-hand panel of Figure A.1 plots the empirical stochastic choice and evaluation functions for *WTP*. Both functions are monotonically increasing, in line with SoP effects in choices and evaluations. As observed previously for the WTA experiments, the stochastic evaluation function is shifted downwards relative to the stochastic choice function, which itself is roughly symmetric around zero. Thus, also in *WTP* DMs exhibit a $-bias in evaluations elicited via WTP valuations. The proportion of P-choices was $\pi_c(WTP) = 0.59$, whereas the proportion of WTP valuations that favored the P-bet was $\pi_v(WTP) = 0.22$. That is, also the *WTP* experiment is biased toward P-bets and

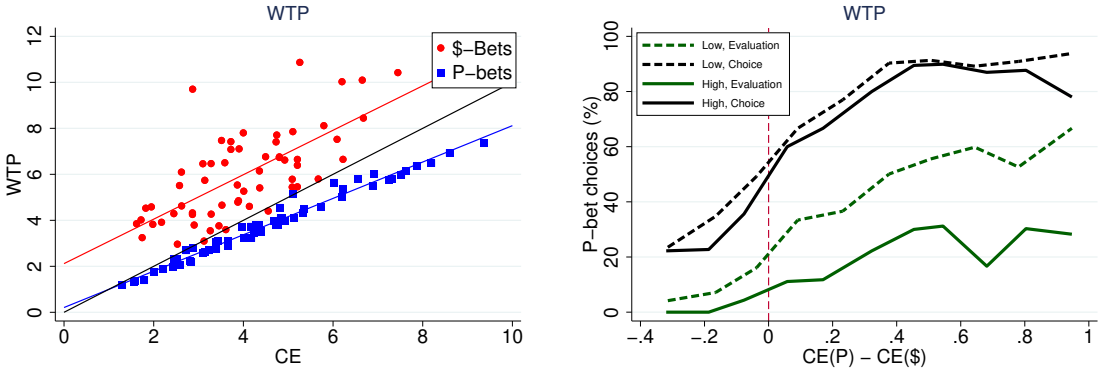


Figure A.2: Accuracy of valuations in *WTP* (left). Median split of β_i (high versus low) for *WTP* (right).

π_c^i is significantly larger than π_v^i (MWW test; $N = 102$, $z = 8.409$, $p < 0.001$). In line with Proposition 3 and despite *WTP* being biased toward P-bets, we find a clear type-2 anomaly with more standard (62.18%) than non-standard reversals (6.27%; WSR test, $N = 94, z = 8.157, p < 0.001$; See Figure A.1, right).

To test Proposition 4, we turn to the accuracy of subjects' WTP valuations given by $\beta_i(\text{P}) = \frac{1}{K}\sum_k \frac{WTP_i(\text{P}_k) - CE_i(\text{P}_k)}{CE_i(\text{P}_k)}$ and $\beta_i(\$) = \frac{1}{K}\sum_k \frac{WTP_i(\$_k) - CE_i(\$_k)}{CE_i(\$_k)}$. Figure A.2 (left) plots the stated WTPs against the estimated CEs for each of the 120 lotteries, distinguishing P-bets and \$-bets. Also WTP-valuations capture subjects' estimated CEs very well for P-bets (Spearman; $\rho = 0.986$, $N = 60$, $p < 0.001$), whereas \$-bets are further away from the diagonal (Spearman; $\rho = 0.692$, $N = 60$, $p < 0.001$). Again, we take $\beta_i = \beta_i(\$) - \beta_i(\text{P})$ as a measure of a DM's \$-bias in WTP-valuations in monetary terms, which we quantify at 63.25% relative to the CE. Figure A.2 (right) shows the empirical stochastic choice and evaluation functions separately for the high and low \$-bias groups (according to a median split on β_i). Again, the high \$-bias group exhibits a stronger \$-bias in evaluations as expected. Comparing reversal rates across groups, we observe that the difference between SR and NR is 68.0 and 42.7 percentage points for the high and low \$-bias groups, respectively. That is, again a stronger \$-bias exacerbates the asymmetry increasing the difference between SR and NR (MWW test; $N = 94$, $z = -3.187$, $p = 0.001$).

Summarizing, our results suggest a \$-bias also in WTP-valuations, which is of a similar magnitude than the one observed in *WTA2*. Overall, the results in experiment *WTP* closely resemble the ones obtained for *WTA2*.

Appendix C Price-based Utility Estimation

The analysis in the main text relied on choice-based, out-of-sample estimates of subject's individual utility functions. In experiments *RANK2* and *WTA2*, we also elicited WTA-

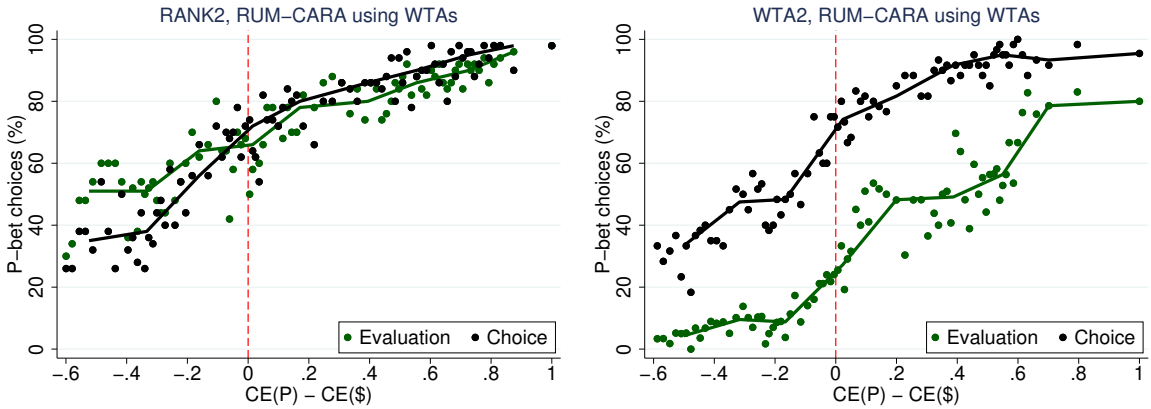


Figure A.3: Strength-of-preference effects in *RANK2* (left) and *WTA2* (right).

valuations for the 64 lotteries used in the first part for the choice-based estimation. Using those WTA-valuations we repeated the estimation exercise described in Appendix Appendix A. Specifically, for each of the 32 lottery pairs (A, B) used in the first part of the experiments, we consider A to be "chosen" by subject i if and only if i stated a higher WTA for A than for B. Using these imputed choices delivers valuation-based utility estimates, which we denote by u_i'. The following analysis proceeds along the lines of Section 5 with the only difference that the certainty equivalent differences $\hat{\Delta}_i(\mathrm{P}_k, \$_k)$ are based on the valuation-based estimates u_i' and not on the choice-based estimates u_i.

Figure A.3 plots the empirical stochastic choice and evaluation functions for *RANK2* (left) and *WTA2* (right). Also with valuation-based CE estimates SoP effects in choices and evaluations are evident in both experiments. As before, in *RANK2* we observe no systematic difference between the stochastic choice and evaluation function, whereas in *WTA2* the latter is clearly shifted downwards relative to the stochastic choice function. Thus, we confirm our previous observation that the ranking-based evaluation method is unbiased, whereas WTA-valuations exhibit a \$-bias.

We obtain the following valuation-based estimates of subjects' risk attitudes. The majority of subjects is risk seeking in both experiments: only 15 subjects or about 12.96% are classified as risk averse in *RANK2*, and 28 subjects or about 27.18% in *WTA2*. Average estimated risk propensities, $\hat{r}$, are -0.023 with a median of -0.024 and a standard deviation of 0.020 in *RANK2* and -0.014 with a median of -0.013 and a standard deviation of 0.027 in *WTA2*.

Proposition 2(a) for unbiased evaluations states that a larger proportion of P-bet choices ($\pi_c(D)$) should lead to a larger type-1 anomaly. A median split of subjects according to the valuation-based estimates of their risk attitudes in *RANK2* does not produce differences in the proportion of P-bet choices (high risk aversion group 55.28%, low risk aversion group 54.72%; MWW; $N = 108$, $z = 0.012$, $p = 0.991$). However, we can directly test the prediction with a median split on $\pi_c(D)$ (average proportion of P-bet choices 39.90% and 67.08% for the below-median and above-median groups, respectively). In the below-median group the rates of standard and non-standard reversals

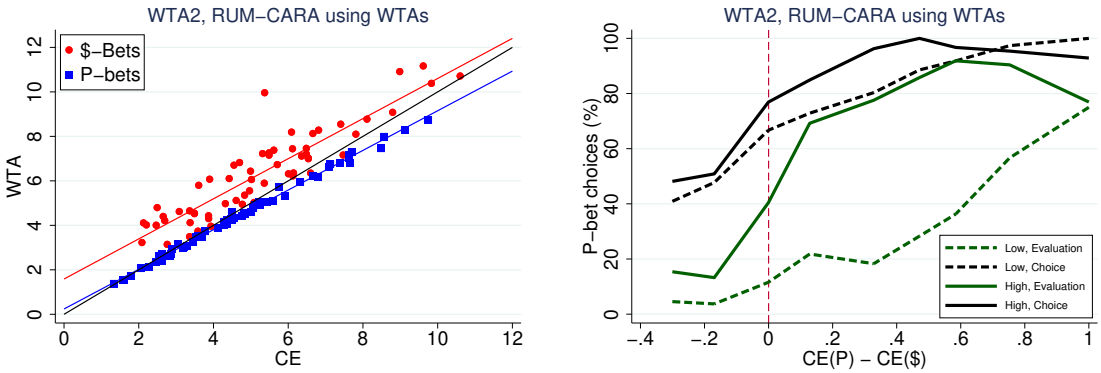


Figure A.4: Accuracy of valuations (left) and median split of β_i (right) for *WTA2*.

are 28.83% and 28.94%, respectively. In contrast, for the above-median group the average rates of standard and non-standard reversals are 16.90% and 46.19%, respectively. The ratio of standard to non-standard reversals is 0.996 for the below-median group and 0.366 for the above-median group. The differences are statistically significantly (MWW test; $N = 104$, $z = -4.176$, $p < 0.001$), in line with Proposition 2(a). That is, a stronger bias in the experiment toward P-bets exacerbates type-1 anomalies in experiments that rely on unbiased evaluation methods.

To test Proposition 4 for biased evaluations, we again resort to β_i(P) and β_i(\$) as a measure of subject i's accuracy in evaluations but now with respect to the valuation-based estimates of subjects' certainty equivalents in *WTA2*. Figure A.4 (left) plots the stated WTAs against the estimated CEs for each of the 120 lottery pairs. For P-bets, the correlation is close to unity (Spearman; *WTA2*: $\rho = 0.997$, $N = 60$, $p < 0.001$) and again the (WTA,CE) pairs are close to the diagonal. For the valuation-based estimates the correlation is also high for \$-bets (Spearman; *WTA2*: $\rho = 0.900$, $N = 60$, $p < 0.001$), however, the (WTA,CE) pairs are again systematically above the diagonal. Thus, even with valuation-based estimates WTAs faithfully reflect CEs for P-bets but not for \$-bets.

We now divide subjects into two groups based on a median split of their \$-bias (based on the u_i' estimates) as captured by β_i. Figure A.4 (right) shows the empirical stochastic choice and evaluation functions for both groups. For the high \$-bias group the stochastic evaluation function is again shifted downwards compared to the low \$-bias group, while the stochastic choice functions are indistinguishable. Comparing the reversal rates across groups, we find a larger asymmetry between standard and non-standard reversals for the high \$-bias group compared to the low \$-bias group (*WTA2*: low, $SR = 46.95\%$, $NR = 7.42\%$; high, $SR = 64.06\%$, $NR = 3.12\%$; MWW test $N = 98$, $z = 3.168$, $p = 0.001$).

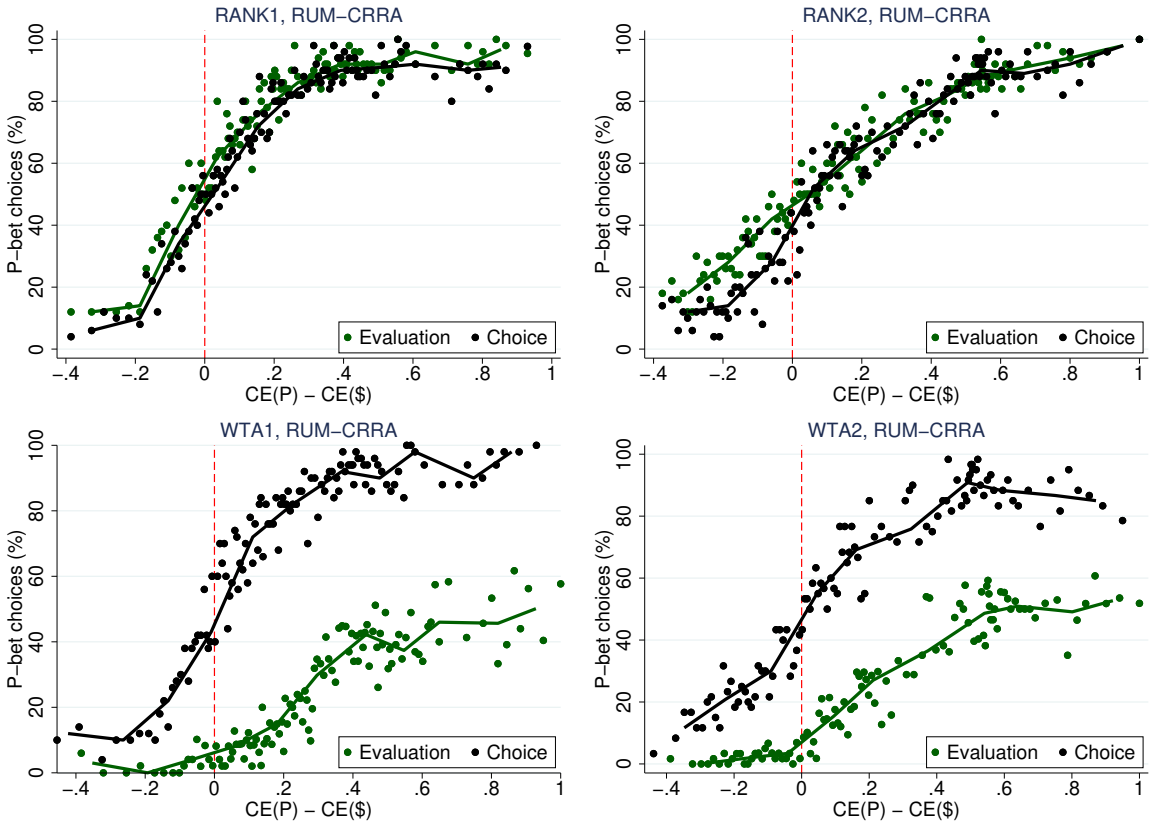


Figure A.5: CRRA. Strength-of-preference effects in *RANK1*, *RANK2*, *WTA1*, and *WTA2*.

Appendix D Alternative Estimation with CRRA Utility

The analysis in the main text relied on utility estimates assuming a CARA utility function. As a further robustness check, we repeated the RUM-based estimation exercise described in Section 4.3 and Appendix Appendix A with the constant relative risk aversion (CRRA) utility function $u(x \mid r) = x^r$. As we detail below, this robustness check confirms that the results reported in Section 5 do not hinge on the CARA specification of the utility function.

Figure A.5 plots the stochastic choice and evaluation functions based on the CRRA-estimates for experiments *RANK1*, *RANK2*, *WTA1*, and *WTA2*. In all four experiments SoP effects are evident. In the ranking-based experiments, both functions are indistinguishable. In contrast, in the WTA experiments we confirm that the stochastic evaluation functions are shifted downwards relative to the stochastic choice functions, which themselves are roughly symmetric around zero. Also based on the CRRA-estimates, the majority of subjects in all four experiments is risk averse (*RANK1*: 98.95%; *RANK2*: 91.67%; *WTA1*: 97.89%; *WTA2*: 99.03%). Conducting a median split based on the estimated CRRA risk parameter, $\hat{r}$, we observe that for the high risk aversion group in *RANK1* (*RANK2*) the proportion of P-bet choices is 75.77% (59.51%), whereas it is

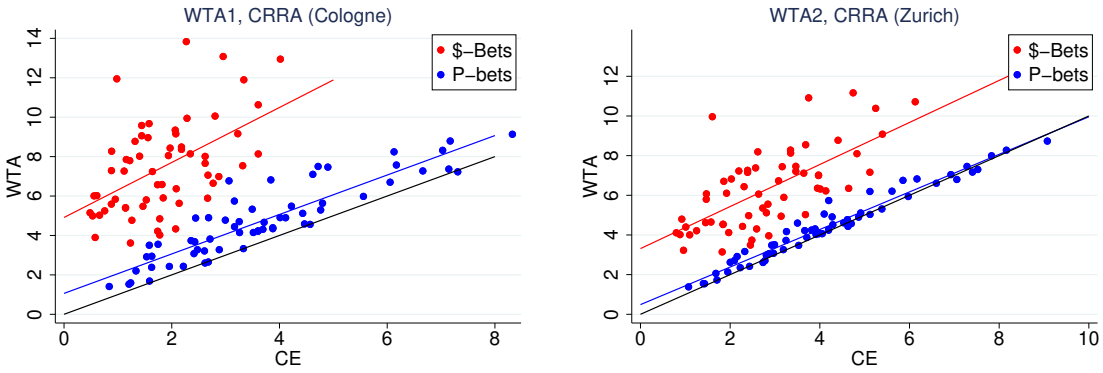


Figure A.6: CRRA. Accuracy of valuations in *WTA1* (left) and *WTA2* (right).

only 55.00% (47.33%) for the low risk aversion group. These differences are significant (MWW tests; *RANK1*: $N = 95$, $z = 5.232$, $p < 0.001$; *RANK2*: $N = 108$, $z = 3.904$, $p < 0.001$). In line with Proposition 2(a), we find that for the high risk aversion group the ratio of standard to non-standard reversals is smaller than for the low risk aversion group (*RANK1*: low, $SR = 9.72\%$, $NR = 39.06\%$; high $SR = 20.72\%$; $NR = 69.31\%$; MWW test, $N = 86$, $z = 4.095$, $p < 0.001$; *RANK2*: low, $SR = 24.64\%$, $NR = 34.08\%$; high, $SR = 19.82\%$, $NR = 43.68\%$; MWW test, $N = 107$, $z = 2.041$, $p = 0.041$).

To test Proposition 4, we again turn to the accuracy of subjects' WTA valuations but now with respect to the CRRA-based estimates of subjects' certainty equivalents. Figure A.6 plots the stated WTAs against the estimated CEs. We again replicate our previous finding that WTAs of P-bets are well-captured by the estimated CEs (Spearman; *WTA1*: $\rho = 0.887$, $N = 60$, $p < 0.001$; *WTA2*: $\rho = 0.972$, $N = 60$, $p < 0.001$), whereas for \$-bets the (WTA,CE) pairs are far away from the diagonal, and the correlation is much lower (Spearman; *WTA1*: $\rho = 0.490$, $N = 60$, $p < 0.001$; *WTA2*: $\rho = 0.646$, $N = 60$, $p < 0.001$). Based on the CRRA estimation, we quantify the \$-bias in *WTA1* and *WTA2* at 320% and 174% relative to the CE, respectively. Next, we divide subjects into two groups based on a median split of their \$-bias, quantified by β_i. Figure A.7 shows the empirical stochastic choice and evaluation functions separately for both groups in *WTA1* and *WTA2*. For the high \$-bias groups, we again find that the stochastic evaluation functions are shifted downwards relative to the low \$-bias group, indicating that the former exhibits a stronger \$-bias in evaluations. Comparing reversal rates, we find, in line with Proposition 2, that the difference between SR and NR in *WTA1* (*WTA2*) is 44.61 (31.18) percentage points for the low \$-bias group, whereas it is 75.55 (67.19) percentage points for the high \$-bias group. These differences are significant (MWW tests, *WTA1*: $N = 86$, $z = -4.986$, $p < 0.001$; *WTA2*: $N = 98$, $z = -6.098$, $p < 0.001$). That is, we again find that a stronger \$-bias exacerbates the asymmetry between standard and non-standard reversals.

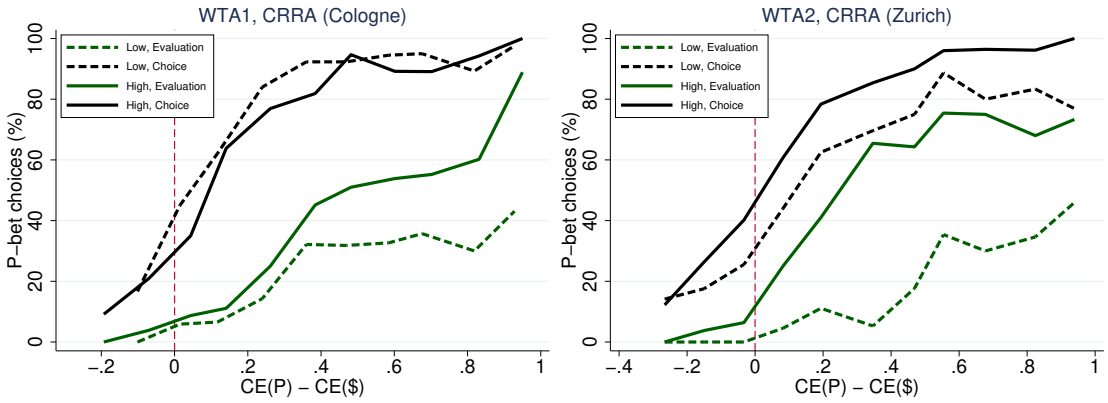


Figure A.7: CRRA. Median split of β_i (high versus low) for *WTA1* (left) and *WTA2* (right).

Appendix E Alternative Estimation with RPM

The analysis in the main text relied on RUM-based estimates. As a further robustness check, we conducted an alternative estimation exercise using an RPM procedure (described below) assuming a CARA utility function. As we outline below, this robustness check confirms that our results do not hinge on the RUM procedure.

Figure A.8 plots the stochastic choice and evaluation functions based on the RPM estimates for *RANK1*, *RANK2*, *WTA1*, and *WTA2*. In all four experiments we observe clear SoP effects. In *RANK1* and *RANK2*, the stochastic choice and evaluation functions are indistinguishable. In contrast, in *WTA1* and *WTA2*, we observe a clear downward shift of the stochastic evaluation functions relative to the stochastic choice functions. The latter ones are roughly symmetric around zero. According to the estimated mean risk parameter $\hat{m}_i$, the majority of subjects in all four experiments is risk averse (*RANK1*: 84.21%; *RANK2*: 52.78%; *WTA1*: 87.37%; *WTA2*: 76.70%). Conducting a median split based on the estimated mean risk parameter, $\hat{m}$, we observe that for the high risk aversion group in *RANK1* (*RANK2*) the proportion of P-bet choices is 80.21% (63.27%), whereas it is only 56.21% (46.73%) for the low risk aversion group. These differences are significant (MWW tests; *RANK1*: $N = 95$, $z = 6.093$, $p < 0.001$; *RANK2*: $N = 108$, $z = 4.942$, $p < 0.001$). We again confirm the prediction of Proposition 2(a). For the high risk aversion group the ratio of standard to non-standard reversals is smaller than for the low risk aversion group (*RANK1*: low, $SR = 22.03\%$, $NR = 30.22\%$; high $SR = 6.83\%$; $NR = 63.60\%$; MWW test, $N = 91$, $z = 6.258$, $p < 0.001$; *RANK2*: low, $SR = 26.63\%$, $NR = 32.40\%$; high, $SR = 17.72\%$, $NR = 44.71\%$; MWW test, $N = 104$, $z = 3.631$, $p = 0.001$).

We again turn to the accuracy of subjects' WTA valuations relative to the RPM-based estimates of subjects' CEs illustrated in Figure A.9. As before, for P-bets the WTA valuations are well-predicted by the estimated CEs (Spearman; *WTA1*: $\rho = 0.927$, $N = 60$, $p < 0.001$; *WTA2*: $\rho = 0.992$, $N = 60$, $p < 0.001$), whereas WTAs for $-bets

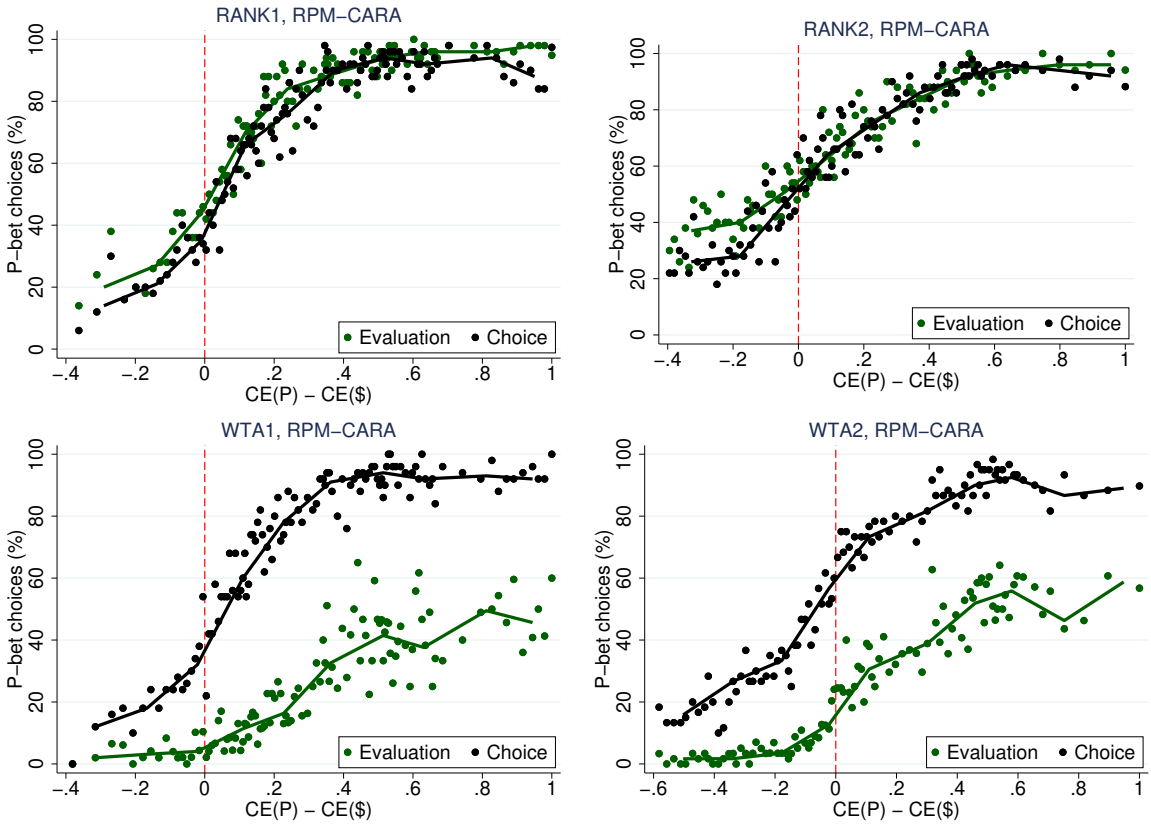


Figure A.8: RPM. Strength-of-preference effects in *RANK1*, *RANK2*, *WTA1*, and *WTA2*.

are further off (Spearman; *WTA1*: $\rho = 0.502$, $N = 60$, $p < 0.001$; *WTA2*: $\rho = 0.764$, $N = 60$, $p < 0.001$). Based on the RPM estimation, we quantify the \$-bias in *WTA1* and *WTA2* at 317% and 60% relative to the CE, respectively. Conducting a median split according to subjects' \$-bias, quantified by β_i, we can again plot the empirical stochastic choice and evaluation functions separately for both groups in *WTA1* and *WTA2* (see Figure A.10). For the high \$-bias groups the stochastic evaluation functions are shifted downwards relative to the low \$-bias groups, that is, the former exhibit a stronger \$-bias in evaluations. We also confirm again Proposition 4. Specifically, the difference between SR and NR in *WTA1* (*WTA2*) is 45.13 (32.23) percentage points for the low \$-bias group, whereas it is 74.26 (66.90) percentage points for the high \$-bias group. These differences are significant (MWW tests, *WTA1*: $N = 86$, $z = -2.326$, $p = 0.019$; *WTA2*: $N = 98$, $z = -5.558$, $p < 0.001$). That is, we again find that a stronger \$-bias exacerbates the asymmetry between standard and non-standard reversals.

Description of the RPM Procedure

For the RPM estimation, we used the same setup with N subjects, $T = 32$ trials, and the CARA utility function given by (6). Additionally, we assume that A_t is the safer of the two lotteries, that is, $p > q$. In contrast to the RUM approach, the RPM assumes

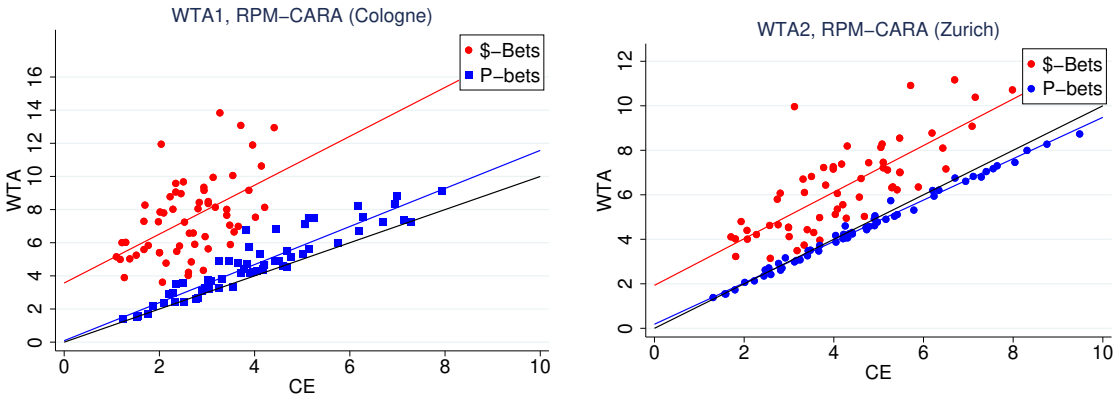


Figure A.9: RPM. Accuracy of valuations in *WTA1* (left) and *WTA2* (right).

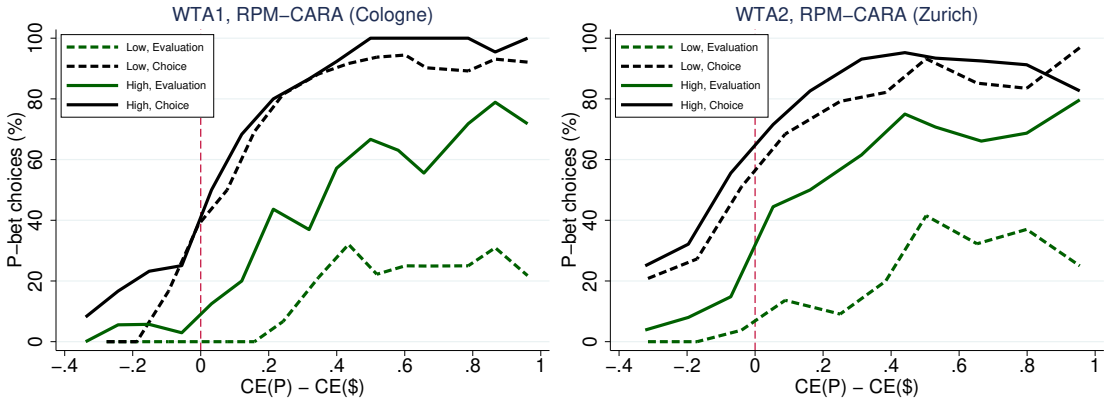


Figure A.10: RPM. Median split of β_i (high versus low) for *WTA1* (left) and *WTA2* (right).

that a subject's risk parameter is not fixed across trials but varies randomly between trials. Specifically, we assume that subject i's risk parameter in trial t is distributed according to $r_{it} \sim N(m_i, \sigma^2)$ where m_i is subject i's mean risk attitude. Assuming Expected Utility maximization, in this setup subject i with utility function u_i chooses A_t over B_t if and only if

$$\Delta_t(r_{it}) = \frac{p_t(1 - e^{-r_{it}x_t}) - q_t(1 - e^{-r_{it}y_t})}{1 - e^{-r_{it}x_{\max}}} > 0.$$

Let r_t^* be the risk parameter that would make a subject exactly indifferent between the two lotteries in task t, that is, $\Delta_t(r_t^*) = 0$. Since A_t is always the safer lottery, we obtain the following equivalence

$$\Delta_t(r_{it}) > 0 \quad \Leftrightarrow \quad r_{it} > r_t^*.$$

Again using $\gamma_{it} \in \{1, -1\}$ as a binary indicator that A_t is chosen by subject i in trial t, the probability of a choice conditional on a subject's mean risk attitude m_i is given by

$$P(\gamma_{it}|m_i) = P(\gamma_{it} r_{it} > \gamma_{it} r_t^* | m_i) = P\left(\gamma_{it} \frac{r_{it} - m_i}{\sigma} > \gamma_{it} \frac{r_t^* - m_i}{\sigma}\right) = \Phi\left(\gamma_{it} \frac{m_i - r_t^*}{\sigma}\right)$$

where Φ is the standard normal cumulative distribution function. Next, in order to introduce between-subject heterogeneity we let the individual mean risk attitude vary across the population. In particular, we assume that

$$m_i \sim N(\mu, \eta^2).$$

Hence, the log-likelihood for a sample consisting of T trials and N subjects given by the matrix $\Gamma = (\gamma_{it})$ is

$$\log L = \sum_{i=1}^{N} \ln \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{m - r_t^*}{\sigma}\right) f(m \mid \mu, \eta) dm \tag{13}$$

where $f(m \mid \mu, \eta)$ is the density function of the mean risk attitude m. Using the MSL approach we replace the integral in (13) by the following approximation

$$\frac{1}{H} \sum_{h=1}^{H} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{m_{ih} - r_t^*}{\sigma}\right)\right) \tag{14}$$

using a sequence of H (transformed) Halton draws $(m_{i1}, \ldots, m_{iH})$ from $N(\mu, \eta^2)$ for each subject i (fixed over trials t). We then maximize the resulting function

$$\log \hat{L} = \sum_{i=1}^{N} \ln \frac{1}{H} \sum_{h=1}^{H} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{m_{ih} - r_t^*}{\sigma}\right)\right). \tag{15}$$

using standard MLE routines in STATA to obtain the parameter estimates $(\hat{\mu}, \hat{\eta}, \hat{\sigma})$. Given those estimates we can then compute the posterior expectation of a subject's mean risk attitude $\hat{m}_i$ conditional on the observed T choices using Bayes' rule as follows

$$\hat{m}_i = E(m_i|\gamma_{i1}, ..., \gamma_{iT}) \approx \frac{\frac{1}{H} \sum_{h=1}^{H} m_{ih} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{m_{ih} - r_t^*}{\sigma}\right)\right)}{\frac{1}{H} \sum_{h=1}^{H} \left(\prod_{t=1}^{T} \Phi\left(\gamma_{it} \frac{m_{ih} - r_t^*}{\sigma}\right)\right)}$$

for a sequence of Halton draws $(m_{i1}, \ldots, m_{iH})$ from $N(\hat{\mu}, \hat{\eta}^2)$.

Given the estimated individual mean risk parameter $\hat{m}_i$, we obtain

$$\hat{u}_i(x) = \frac{1 - e^{-\hat{m}_i x}}{1 - e^{-\hat{m}_i x_{\max}}} \text{ for } \hat{m}_i \neq 0$$

as the estimated utility function of subject i.

Appendix F List of Lottery Pairs

Table A.1 shows the 32 lottery pairs used for the utility estimations (first part). Table A.2 shows the 60 lottery pairs used in the preference reversal experiments (second part).

Table A.1: Lottery pairs (A, B) used for the utility estimation in the first part.

Lottery	Lottery A			Lottery B		
Pair	p	x	EV	q	y	EV
1	0.05	12	0.6	0.8	3	2.4
2	0.2	22	4.4	0.8	5	4
3	0.25	17	4.25	0.75	6	4.5
4	0.35	20	7	0.6	8	4.8
5	0.35	17	5.95	0.7	4	2.8
6	0.4	12	4.8	0.7	6	4.2
7	0.4	14	5.6	0.65	6	3.9
8	0.4	14	5.6	0.8	3	2.4
9	0.5	11	5.5	0.7	7	4.9
10	0.5	15	7.5	0.65	7	4.55
11	0.5	20	10	0.7	5	3.5
12	0.55	5	2.75	0.35	18	6.3
13	0.55	4	2.2	0.2	15	3
14	0.55	4	2.2	0.4	15	6
15	0.55	4	2.2	0.45	21	9.45
16	0.6	6	3.6	0.35	11	3.85
17	0.6	5	3	0.3	22	6.6
18	0.6	8	4.8	0.5	13	6.5
19	0.6	14	8.4	0.7	4	2.8
20	0.6	4	2.4	0.55	6	3.3
21	0.6	3	1.8	0.5	13	6.5
22	0.65	3	1.95	0.15	18	2.7
23	0.65	17	11.05	0.75	7	5.25
24	0.7	4	2.8	0.1	16	1.6
25	0.7	7	4.9	0.6	20	12
26	0.7	11	7.7	0.8	6	4.8
27	0.7	18	12.6	0.85	5	4.25
28	0.75	6	4.5	0.3	15	4.5
29	0.75	6	4.5	0.4	15	6
30	0.75	4	3	0.35	12	4.2
31	0.75	15	11.25	0.8	5	4
32	0.8	3	2.4	0.4	17	6.8

Table A.2: Lottery pairs (P, \$) used in the preference reversal experiments.

Lottery	P-Bets			\$-Bets		
Pair	p	x	EV	q	y	EV
1	0.95	3	2.85	0.37	10	3.70
2	0.57	5	2.85	0.46	10	4.60
3	0.90	6	5.40	0.30	11	3.30
4	0.80	6	4.80	0.30	11	3.30
5	0.72	7	5.04	0.23	11	2.53
6	0.79	2	1.58	0.21	11	2.31
7	0.8	2	1.60	0.4	11	4.40
8	0.64	8	5.12	0.24	12	2.88
9	0.84	6	5.04	0.48	12	5.76
10	0.75	3	2.25	0.17	12	2.04
11	0.94	3	2.82	0.49	12	5.88
12	0.92	4	3.68	0.53	12	6.36
13	0.82	3	2.46	0.34	12	4.08
14	0.74	6	4.44	0.15	13	1.95
15	0.89	5	4.45	0.39	13	5.07
16	0.87	6	5.22	0.36	13	4.68
17	0.9	2	1.80	0.35	13	4.55
18	0.66	2	1.32	0.24	13	3.12
19	0.6	5	3.00	0.45	13	5.85
20	0.9	7	6.30	0.51	14	7.14
21	0.86	5	4.30	0.16	15	2.40
22	0.70	10	7.00	0.31	15	4.65
23	0.85	5	4.25	0.41	15	6.15
24	0.63	7	4.41	0.41	15	6.15
25	0.75	6	4.50	0.15	15	2.25
26	0.76	11	8.36	0.37	16	5.92
27	0.63	4	2.52	0.33	16	5.28
28	0.96	5	4.80	0.19	17	3.23
29	0.96	8	7.68	0.43	17	7.31
30	0.84	9	7.56	0.25	18	4.50
31	0.83	6	4.98	0.31	18	5.58
32	0.95	5	4.75	0.22	18	3.96
33	0.86	5	4.30	0.33	18	5.94
34	0.79	4	3.16	0.33	18	5.94
35	0.60	11	6.60	0.22	19	4.18
36	0.56	10	5.60	0.43	19	8.17
37	0.79	7	5.53	0.20	20	4.00
38	0.7	5	3.50	0.17	20	3.40
39	0.85	10	8.50	0.3	20	6.00
40	0.65	4	2.60	0.25	20	5.00
41	0.92	8	7.36	0.23	21	4.83
42	0.88	11	9.68	0.35	21	7.35
43	0.72	6	4.32	0.29	21	6.09
44	0.68	3	2.04	0.23	21	4.83
45	0.73	9	6.57	0.21	22	4.62
46	0.6	7	4.20	0.3	22	6.60
47	0.68	11	7.48	0.23	23	5.29
48	0.88	8	7.04	0.4	24	9.60
49	0.84	7	5.88	0.35	25	8.75
50	0.95	8	7.60	0.31	27	8.37
51	0.82	11	9.02	0.24	31	7.44
52	0.87	5	4.35	0.13	32	4.16
53	0.86	4	3.44	0.55	6	3.30
54	0.8	4	3.20	0.45	6	2.70
55	0.87	3	2.61	0.5	7	3.50
56	0.75	5	3.75	0.55	7	3.85
57	0.82	5	4.10	0.47	8	3.76
58	0.71	5	3.55	0.22	9	1.98
59	0.89	5	4.45	0.55	9	4.95
60	0.82	4	3.28	0.36	9	3.24

Appendix G Translated Instructions

[*These are the written instructions for RANK1 and WTA1 given to subjects before the experiment. The original instructions for the Cologne experiments were in German. Text in brackets [...] was not displayed to subjects. The instructions for the other experiments were very similar and are available from the authors upon request.*]

General Instructions

Welcome! In this experiment you will be asked to make a series of decisions that will determine your earnings at the end of the experiment. The total duration of the experiment is about 1 hour and 30 minutes.

If you have a question, please raise your hand and remain seated. An experimenter will come and answer your question.

It is important, that you read the instructions carefully before you make your decisions.

During the experiment you are not allowed to talk or communicate in any other way with the other participants. If you violate this rule, you might be excluded from the experiment.

We now explain the general course of the experiment: The experiment consists of three parts. In each part you have to make multiple decisions. At the end of the experiment you will be asked to answer a questionnaire.

In each part, you can earn money. How much money you earn will depend on your decisions in that part and chance. Your earnings in one part of the experiment are independent of your earnings and decisions in the other parts. Your earnings in each part will be added up and you will be paid the total amount anonymously and in cash at the end of the experiment. In addition to this amount you will receive €4 for your participation in the experiment.

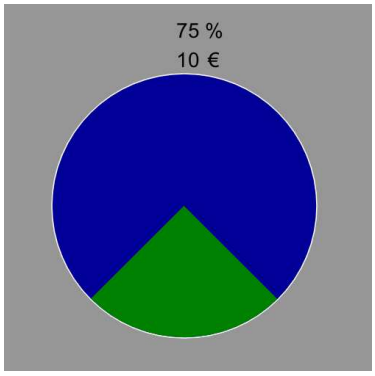
Below you will find further general information for the experiment. The specific instructions for each part will be shown on screen directly before the beginning of that part.

Instructions: Lotteries

In the three parts of the experiment you will be asked to make decisions about lotteries. Hence, we will now explain in detail what a lottery is:

A lottery consists of two potential outcomes, each of which will occur with a given probability. One of the two outcomes is always €0 (zero). The other outcome will differ from lottery to lottery. If a lottery is played out, this means that you will receive exactly one of the two possible outcomes (in Euro).

In the experiment lotteries will be represented by pie charts as in the example below. The colored areas of the pie chart illustrate the probabilities for the two corresponding outcomes.



Example:

The pie chart depicted above is an example of how we present a lottery. In this example, the lottery pays €10 with a probability of 75%, which is represented by the blue area. Additionally, this information is also shown numerically above the pie chart. Accordingly, the lottery pays €0 with a probability of 25%, which is represented by the green area. The second outcome is always €0 and occurs with the remaining probability. Please note that this information is not repeated numerically on screen.

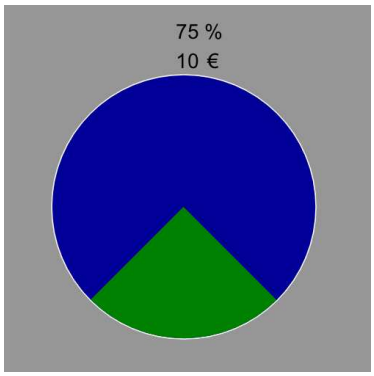
If a lottery is played out, this means that it will pay exactly one of the two outcomes. In the example above, the lottery pays €10 with a probability of 75% and €0 with the remaining probability of 25%.

Please note that the lottery shown above is only an example. The lotteries in the experiment will have different outcomes and probabilities.

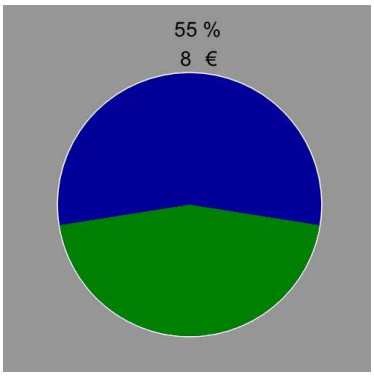
If you have a question, please raise your hand. If you have no further questions, you may proceed to the comprehension questions on the next page.

Comprehension Questions: Lotteries

Below you see examples of two lotteries, similar to the ones you will face later on in the experiment. Please note that these lotteries are only examples.



Lottery A



Lottery B

Please answer the following comprehension questions:

1. What is the probability that Lottery A pays €10?
2. What is the probability that Lottery B pays €0?
3. Which amount does Lottery A pay with a probability of 25%?
4. Which amount does Lottery B pay with a probability of 55%?

Once you have answered all comprehension questions, please raise your hand. An experimenter will then check your answers.

Translated Onscreen Instructions

[These are the instructions for each part, which were presented separately on screen, at the beginning of each part. The original instructions were in German. Text in brackets [...] was not displayed to subjects.]

Welcome to this economic experiment. Thank you for supporting our research.

Please note the following rules:

1. During the experiment you are not allowed to communicate with each other.
2. If you have questions, please raise your hand.
3. Please refrain from using any features of the computer that are not part of the experiment.

Instructions for Part 1

Your decisions: In this part of the experiment you will be presented with a series of lottery pairs. Your task is to choose one of the two lotteries from each pair.

On the screen you will see a lottery pair (consisting of two lotteries) represented by two pie charts. One of the lotteries will be shown on the left and the other will be shown on the right. You choose one of the lotteries by pressing the left or right arrow key on your keyboard. These keys are marked with a yellow sticker. To choose the lottery on the left, press the left arrow key "←." To choose the lottery on the right, press the right arrow key "→." Please note that your decisions will affect your earnings at the end of the experiment (a detailed description of how your earnings are determined will follow below).

There are no wrong or correct decisions. When you choose one of the lotteries, this simply shows that you prefer to play this lottery over the other lottery.

After you have made your decision, you will see the next lottery pair. In part 1 you will be presented with a total of 36 lottery pairs. After you have made a decision for each of the pairs, this part ends and we will start with the next part of the experiment.

Your Earnings for Part 1

After you have made a decision for each of the lottery pairs, the computer will randomly select one of the 36 lottery pairs. The computer then checks which of the two lotteries you have chosen for this randomly selected pair. The lottery you have chosen will be played out. The outcome of the lottery determines your earnings for part 1 of the experiment.

The lottery will be played out at the end of the experiment, that is, after you have completed all three parts of the experiment. Please note: Although your earnings for this part will be determined at the end of the experiment, they will only depend on your decisions in this part of the experiment and chance.

If you have any further questions, please raise your hand and remain seated.

Instructions for Part 2 [Experiment *WTA1*]

Your decisions: In this part of the experiment you will be presented with a series of lotteries. When a lottery is presented to you on screen, you may simply assume, that you own that lottery and are asked to sell it.

Your task is to state the lowest price at which you are still willing to sell the presented lottery instead of keeping the lottery and playing it out. There is no wrong or correct answer when stating the lowest price at which you are still willing to sell the lottery. When you enter your selling price for the lottery, simply ask yourself "Is this really the lowest price at which I am still willing to sell the lottery instead of playing the lottery?". Please note that your decisions will affect your earnings at the end of the experiment (a detailed description of how your earnings are determined will follow below).

Please enter the lowest price at which you are still willing to sell the lottery in the form "EURO.CENTS." Please note that you cannot enter a selling price that is larger than the highest outcome of the lottery.

After you have entered your selling price, the next lottery will be presented. In this part of the experiment you will see a total of 120 lotteries, presented in 20 rounds of 6 lotteries each. All rounds are independent. Once you have entered a selling price for each lottery in a round, the next round will start. Once you are done with all 20 rounds, you can continue with the next part of the experiment.

Your Earnings for Part 2 [Experiment *WTA1*]

After you have entered your lowest selling price for each of the lotteries, the computer will randomly draw one of the 20 rounds. From this round, the computer will then randomly select two of the six lotteries. The computer then checks for which of the two lotteries you have entered the higher selling price (in case both prices are the same, the computer will randomly select one of the two lotteries with equal probability). This lottery will be played out and the outcome of that lottery determines your earnings for part 2 of the experiment.

The lottery will be played out at the end of the experiment, that is, after you have completed all three parts of the experiment. Please note: Although your earnings for this part will be determined at the end of the experiment, they will only depend on your decisions in this part of the experiment and chance.

If you have any further questions, please raise your hand and remain seated.

Instructions for Part 2 [Experiment *RANK1*]

Your decisions: In this part of the experiment you will be presented with a series of lotteries. When a lottery is presented to you on screen, you may simply assume, that you own that lottery and may play that lottery.

Your task is to order different lotteries according to your preference, that is, according to how much you would like to play them. In each round you will see six different lotteries on screen. Please order the lotteries as follows:

- First, choose your first-ranked lottery, that is, the one of the six lotteries that you would like to play the most.
- Second, choose your second-ranked lottery, that is the one that you would like to play out the second most.
- Third, choose your third-ranked lottery, that is the one that you would like to play out the third most.
- Fourth, choose your fourth-ranked lottery, that is the one that you would like to play out the fourth most.
- Fifth, choose your fifth-ranked lottery, that is the one that you would like to play out the fifth most.
- Sixth, choose your sixth-ranked lottery, that is the one that you would like to play out the least.

To select a lottery simply click on the button below the lottery that you want to select. As soon as you assign a rank to a lottery, the corresponding rank (from 1 to 6) will be shown below that lottery.

In case you want to change the rank of the lotteries, please press the "Reset" button. This resets the ranking. After you have ranked the lotteries from rank 1 to rank 6, please press the "Continue" button to confirm your ranking and proceed to the next round.

Please note that there is no wrong or correct ranking. When ranking the lotteries, simply ask yourself which lottery you would like to play out the most, the second-most and so on. Please note that your decisions will affect your earnings at the end of the experiment (a detailed description of how your earnings are determined will follow below).

In this part of the experiment you will see a total of 120 lotteries, presented in 20 rounds of 6 lotteries each. All rounds are independent, that is, you will have to submit 20 rankings of 6 lotteries by assigning ranks from 1 to 6. Once you are done with all 20 rounds, you can continue with the next part of the experiment.

Your Earnings for Part 2 [Experiment *RANK1*]

After you have ranked all lotteries, the computer will randomly draw one of the 20 rounds. From this round, the computer will then randomly select two of the six lotteries. The computer then checks which of the two lotteries you have ranked higher (that is, which one you want to play out more). This lottery will be played out and the outcome of that lottery determines your earnings for part 2 of the experiment.

The lottery will be played out at the end of the experiment, that is, after you have completed all three parts of the experiment. Please note: Although your earnings for this part will be determined at the end of the experiment, they will only depend on your decisions in this part of the experiment and chance.

If you have any further questions, please raise your hand and remain seated.

Instructions for Part 3

Your decisions: In this part of the experiment you will be presented with a series of lottery pairs. Similarly to part 1, your task is to choose one of the two lotteries from each pair. Please note that the lottery pairs are different from part 1.

On the screen you will see a lottery pair (consisting of two lotteries) represented by two pie charts. One of the lotteries will be shown on the left and the other will be shown on the right. You can choose one of the lotteries pressing the left or right arrow key on your keyboard. These keys are marked with a yellow sticker. To choose the lottery on the left, press the left arrow key "←." To choose the lottery on the right, press the right arrow key "→." Please note that your decisions will affect your earnings at the end of the experiment (a detailed description of how your earnings are determined will follow below).

There are no wrong or correct decisions. When you choose one of the lotteries, this simply shows that you prefer to play this lottery over the other lottery.

After you have made your decision, you will see the next lottery pair. In part 3 you will be presented with a total of 60 lottery pairs. After you have made a decision for each of the pairs, this part ends and you can start the questionnaire.

Your Earnings for Part 3

After you have made a decision for each of the lottery pairs, the computer will randomly select one of the 60 lottery pairs. The computer then checks which of the two lotteries you have chosen for this randomly selected pair. The lottery you have chosen will be

played out. The outcome of the lottery determines your earnings for part 3 of the experiment.

The lottery will be played out at the end of the experiment, that is, after you have completed all three parts of the experiment. Please note: Although your earnings for this part will be determined at the end of the experiment, they will only depend on your decisions in this part of the experiment and chance.

If you have any further questions, please raise your hand and remain seated.

References

Alós-Ferrer, C. and M. Garagnani (2018). Strength of Preference and Decisions Under Risk. Working Paper, University of Zurich.

Alós-Ferrer, C., D.-G. Granić, J. Kern, and A. K. Wagner (2016). Preference Reversals: Time and Again. *Journal of Risk and Uncertainty 52*(1), 65–97.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2006). Elicitation Using Multiple Price List Formats. *Experimental Economics 9*(4), 383–405.

Apesteguía, J. and M. A. Ballester (2018). Monotone Stochastic Choice Models: The Case of Risk and Time Preferences. *Journal of Political Economy 126*(1), 74–106.

Atkinson, A. C. (1996). The Usefulness of Optimum Experimental Designs. *Journal of the Royal Statistical Society 51*(1), 59–76.

Attema, A. E. and W. B. Brouwer (2013). In Search of a Preferred Preference Elicitation Method: A Test of the Internal Consistency of Choice and Matching Tasks. *Journal of Economic Psychology 39*, 126–140.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018). Incentives in Experiments: A Theoretical Analysis. *Journal of Political Economy 126*(4), 1472–1503.

Bateman, I., B. Day, G. Loomes, and R. Sugden (2007). Can Ranking Techniques Elicit Robust Values? *Journal of Risk and Uncertainty 34*(1), 49–66.

Bateman, I. J., R. T. Carson, B. Day, M. Hanemann, N. Hanley, T. Hett, M. J. Lee, G. Loomes, S. Mourato, E. Ozdemiroglu, D. W. Pearce, R. Sugden, and J. Swanson (2002). *Economic Valuation with Stated Preference Techniques: A Manual*. Cheltenham, United Kingdom: Edward Elgar.

Beauchamp, J. P., D. J. Benjamin, D. I. Laibson, and C. F. Chabris (2019). Measuring and Controlling for the Compromise Effect when Estimating Risk Preference Parameters. *Experimental Economics forthcoming*.

Becker, G. M., M. H. DeGroot, and J. Marshak (1964). Measuring Utility by a Single-Response Sequential Method. *Behavioral Science 9*(3), 226–232.

Bellemare, C., S. Kröger, and A. van Soest (2008). Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities. *Econometrica 76*(4), 815–839.

Bleichrodt, H. and J. L. Pinto Prades (2009). New Evidence of Preference Reversals in Health Utility Measurement. *Health Economics 18*(6), 713–726.

Bruner, D. M. (2017). Does Decision Error Decrease with Risk Aversion? *Experimental Economics 20*(1), 259–273.

Butler, D. J. and G. Loomes (2007). Imprecision as an Account of the Preference Reversal Phenomenon. *American Economic Review 97*(1), 277–297.

Cappellari, L. and S. P. Jenkins (2003). Multivariate Probit Regression Using Simulated Maximum Likelihood. *The Stata Journal 3*(3), 278–294.

Casey, J. T. (1991). Reversal of the Preference Reversal Phenomenon. *Organizational Behavior and Human Decision Processes 48*(2), 224–251.

Chai, X. (2005). Cognitive Preference Reversal or Market Price Reversal? *Kyklos 58*(2), 177–194.

Chu, Y.-P. and R.-L. Chu (1990). The Subsidence of Preference Reversals in Simplified and Marketlike Experimental Settings: A Note. *American Economic Review 80*(4), 902–911.

Collins, S. M. and D. James (2015). Response Mode and Stochastic Choice Together Explain Preference Reversals. *Quantitative Economics 6*(3), 825–856.

Conte, A., J. D. Hey, and P. G. Moffatt (2011). Mixture Models of Choice Under Risk. *Journal of Econometrics 162*(1), 79–88.

Cubitt, R. P., A. Munro, and C. Starmer (2004). Testing Explanations of Preference Reversal. *Economic Journal 114*(497), 709–726.

Dashiell, J. F. (1937). Affective Value-Distances as a Determinant of Aesthetic Judgment-Times. *American Journal of Psychology 50*, 57–67.

Delquié, P. (1993). Inconsistent Trade-Offs Between Attributes: New Evidence in Preference Assessment Biases. *Management Science 39*(11), 1382–1395.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics 10*(2), 171–178.

Ford, I., B. Torsney, and C. J. Wu (1992). The Use of a Canonical Form in the Construction of Locally Optimal Designs for Non-Linear Problems. *Journal of the Royal Statistical Society 54*(2), 569–583.

Goldstein, W. M. and H. J. Einhorn (1987). Expression Theory and the Preference Reversal Phenomena. *Psychological Review 94*(2), 236–254.

Grether, D. M. and C. R. Plott (1979). Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review 69*(4), 623–638.

Halton, J. H. (1960). On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals. *Numerische Mathematik 2*(1), 84–90.

Hershey, J. C. and P. J. H. Schoemaker (1985). Probability versus Certainty Equivalence Methods in Utility Measurement: Are They Equivalent? *Management Science 31*(10), 1213–1231.

Holt, C. A. (1986). Preference Reversals and the Independence Axiom. *American Economic Review 76*(3), 508–515.

Holt, C. A. and S. K. Laury (2002). Risk Aversion and Incentive Effects. *American Economic Review 92*(5), 1644–1655.

Johnson, E. J. and D. A. Schkade (1989). Bias in Utility Assessments: Further Evidence and Explanations. *Management Science 35*(4), 406–424.

Karni, E. and Z. Safra (1987). 'Preference Reversal' and the Observability of Preferences by Experimental Methods. *Econometrica 55*(3), 675–685.

Lichtenstein, S. and P. Slovic (1971). Reversals of Preference Between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology 89*(1), 46–55.

Lichtenstein, S. and P. Slovic (1973). Response-Induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas. *Journal of Experimental Psychology 101*(1), 16–20.

Lindman, H. R. (1971). Inconsistent Preferences Among Gambles. *Journal of Experimental Psychology 89*(2), 390–397.

Lipkus, I. M., G. Samsa, and B. K. Rimer (2001). General Performance on a Numeracy Scale Among Highly Educated Samples. *Medical Decision Making 21*(1), 37–44.

Loomes, G. (2005). Modelling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data. *Experimental Economics 8*(4), 301–323.

Loomes, G. and R. Sugden (1995). Incorporating a Stochastic Element into Decision Theories. *European Economic Review 39*(3–4), 641–648.

Loomes, G. and R. Sugden (1998). Testing Different Stochastic Specifications of Risky Choice. *Economica 65*(260), 581–598.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Maafi, H. (2011). Preference Reversals Under Ambiguity. *Management Science 57*(11), 2054–2066.

McFadden, D. L. (2001). Economic Choices. *American Economic Review 91*(3), 351–378.

Moffatt, P. G. (2005). Stochastic Choice and the Allocation of Cognitive Effort. *Experimental Economics 8*(4), 369–388.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for Experimental Economics*. London: Palgrave Macmillan.

Moyer, R. S. and T. K. Landauer (1967). Time Required for Judgements of Numerical Inequality. *Nature 215*(5109), 1519–1520.

Oliver, A. (2013). Testing Procedural Invariance in the Context of Health. *Health Economics 22*(3), 272–288.

Peirce, J. W. (2007). PsychoPy – Psychophysics Software in Python. *Journal of Neuroscience Methods 162*(1), 8–13.

Pommerehne, W. W., F. Schneider, and P. Zweifel (1982). Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination. *American Economic Review 72*(3), 569–574.

Reilly, R. J. (1982). Preference Reversal: Further Evidence and Some Suggested Modifications in Experimental Design. *American Economic Review 72*(3), 576–584.

Safra, Z., U. Segal, and A. Spivak (1990). Preference Reversal and Nonexpected Utility Behavior. *American Economic Review 80*(4), 922–930.

Schkade, D. A. and E. J. Johnson (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes 44*(2), 203–231.

Schmidt, U. and J. D. Hey (2004). Are Preference Reversals Errors? An Experimental Investigation. *Journal of Risk and Uncertainty 29*(3), 207–218.

Seidl, C. (2002). Preference Reversal. *Journal of Economic Surveys 16*(5), 621–655.

Selten, R., A. Sadrieh, and K. Abbink (1999). Money Does Not Induce Risk Neutral Behavior, but Binary Lotteries Do Even Worse. *Theory and Decision 46*(3), 213–252.

Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation*, Volume 1. New York: Chapman and Hall.

Slovic, P. and S. Lichtenstein (1968). Relative Importance of Probabilities and Payoffs in Risk Taking. *Journal of Experimental Psychology Monograph 78*(3, Part 2), 1–18.

Stalmeier, P. F. M., P. P. Wakker, and T. G. G. Bezembinder (1997). Preference Reversals: Violations of Unidimensional Procedure Invariance. *Journal of Experimental Psychology: Human Perception and Performance 23*(4), 1196–1205.

Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review 34*, 273–286.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

Trautmann, S. T., F. M. Vieider, and P. P. Wakker (2011). Preference Reversals for Ambiguity Aversion. *Management Science 57*(7), 1320–1333.

Tversky, A., S. Sattath, and P. Slovic (1988). Contingent Weighting in Judgment and Choice. *Psychological Review 95*(3), 371–384.

Tversky, A., P. Slovic, and D. Kahneman (1990). The Causes of Preference Reversal. *American Economic Review 80*(1), 204–217.

Tversky, A. and R. H. Thaler (1990). Anomalies: Preference Reversals. *Journal of Economic Perspectives 4*(2), 201–211.

Vieider, F. M. (2018). Violence and Risk Preference: Experimental Evidence from Afghanistan, Comment. *American Economic Review 108*(8), 2366–2382.

Wilcox, N. T. (2008). Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison. In J. C. Cox and G. W. Harrison (Eds.), *Risk Aversion in Experiments*, Volume 12 of *Research in Experimental Economics*, pp. 197–292. Bingley, UK: Emerald.

Wilcox, N. T. (2011). Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk. *Journal of Econometrics 162*(1), 89–104.